Properties of Mean Shift

Ryoya Yamasaki, and Toshiyuki Tanaka, *Member, IEEE*

Abstract—We study properties of the mean shift (MS)-type algorithms for estimating modes of probability density functions (PDFs), via regarding these algorithms as gradient ascent on estimated PDFs with adaptive step sizes. We rigorously prove convergence of mode estimate sequences generated by the MS-type algorithms, under the assumption that an analytic kernel function is used. Moreover, our analysis on the MS function finds several new properties of mode estimate sequences and corresponding density estimate sequences, including the result that in the MS-type algorithm using a Gaussian kernel the density estimate monotonically increases between two consecutive mode estimates. This implies that, in the one-dimensional case, the mode estimate sequence monotonically converges to the stationary point nearest to an initial point without jumping over any stationary point.

Index Terms—Mode estimation, mode clustering, mean shift algorithm, conditional mean shift algorithm, subspace constrained mean shift algorithm

1 Introduction

The mean shift (MS) algorithm [1], [2], [3] is a method to estimate modes of a probability density function (PDF) via gradient ascent of an estimated PDF with adaptive step sizes. The MS algorithm is mainly applied to mode clustering [4], [5]. The MS-based mode clustering is a flexible method that does not require specifying the number of clusters and the initial points of cluster centers beforehand, and it can cope with arbitrary cluster shapes. Also, in the fields of computer vision, image processing, and pattern recognition, the MS algorithm is widely used, for example, in image segmentation [3], [6], edge detection [7], [8], object tracking [9], [10], and so on.

The conditional mean shift (CMS) algorithm [11], [12], [13], a variant of the MS algorithm, is a representative technique for nonparametric modal regression, which has been applied to analysis of traffic data [12], [13] and weather data [14]. The CMS algorithm can be regarded as a weighted version of the conventional MS algorithm with the weights determined by the values of the independent variables in the samples, and it estimates modes of a conditional PDF of the dependent variables conditioned on the independent variables.

Another variant of the MS algorithm is the subspace constrained mean shift (SCMS) algorithm [15], [16], [17], which is a method of estimating principal curves and principal surfaces as ridges of an estimated PDF [18]. It has been applied to face alignment [19] and analysis of cosmic data [20]. This method performs gradient ascent with an adaptive step size similar to the MS algorithm in a suitably constrained subspace at each iteration. In this paper, we use the term "the MS-type algorithms" to collectively refer to these algorithms derived on the basis of the MS algorithm.

• *Ryoya Yamasaki and Toshiyuki Tanaka are with the Department of Systems Science, Graduate School of Informatics, Kyoto University, Kyoto, 606-8501, Japan. E-mail: yamasaki@sys.i.kyoto-u.ac.jp, tt@i.kyoto-u.ac.jp*

TABLE 1
Classification of the main results according to problem dimension and kernel function.

Dimension	Kernel	Results
General	General	Theorem 1 Corollary 1 Propositions 1 & 3 Lemmas 1, 2, & 4 Algorithm 2
	Analytic	Theorem 2 Corollary 2 Propositions 2
	Gaussian	Theorem 3 Propositions 3 Lemma 3
1-dim.	General	Proposition 4
	Gaussian	Theorem 4 Algorithm 3

It has empirically been observed that mode estimate sequences obtained by the MS algorithms efficiently converge to local modes of estimated PDFs. However, our understanding on theoretical properties of these algorithms, such as their convergence properties, is quite limited. In this paper, we study properties of mode estimate sequences generated by the MS-type algorithms and the corresponding density estimate sequences. Then, we give new findings and unified understanding on properties and convergence of the MS-type algorithms (see Table 1).

Although there are two well-known proofs for convergence of the MS algorithm for estimated PDFs given by kernel density estimation [3], [21], it has been pointed out that neither is strictly rigorous. It was claimed in [3] that the mode estimate sequence generated by the MS algorithm is a Cauchy sequence and hence converges. As pointed out in [22], however, there was a flaw in [3] in proving that the mode estimate sequence satisfies the condition of a Cauchy sequence. Convergence of the mode estimate sequence generated by the MS algorithm was also claimed in [21] under the assumption that a Gaussian kernel is used, on the ground that the MS algorithm using a Gaussian

kernel is an instance of the expectation maximization (EM) algorithm. However, as pointed out in [23], the claim is not justifiable since it is known that the EM algorithm may not converge without additional conditions [24]. Also, a recent paper [25], under the condition that sufficiently small step sizes are used, gives an error bound between a piecewise linear trajectory connecting the mode estimate sequence (a mode estimate path in our Definition 5) and a trajectory of the gradient flow of the underlying PDF, and discusses the condition under which the mode estimate sequence converges to proper modes (see our Definition 6). However, their proof [25, middle of p. 15] on convergence of the mode estimate sequence has the same flaw as [3], and consequently, convergence of the mode estimate sequence has not been proven in [25]. We also note that a step size considered in [25] is smaller than the adaptive step size currently used in most applications of the MS algorithm, and consequently, the problem setting in [25] differs from that in other papers discussing the convergence of the MS algorithm.

Several studies attempt to rigorously prove convergence of mode estimate sequences generated by the MS algorithm. In [22], convergence of mode estimate sequences has been proved under the assumption that an estimated PDF has a finite number of stationary points inside the convex hull of samples. It has not been proved, however, whether this assumption holds for estimated PDFs with commonly used kernel functions such as a Gaussian kernel (see e.g., [26]). After that, a sufficient condition that an estimated PDF with a Gaussian kernel satisfies this assumption has been given in [27]. The obtained condition requires taking the scale parameter (bandwidth) of the Gaussian kernel large enough. Under this condition, however, an estimated PDF, as well as the mode estimate sequences obtained therefrom, would have a large bias, making practical significance of the condition quite obscure. Consequently, as far as the authors' knowledge, there has been no complete proof of convergence of the mode estimate sequence generated by the MS algorithm under the multivariate setting as well as with step sizes commonly used in applications. We would like to mention, however, that it has rigorously been proved in [23] that mode estimate sequences generated by the MS algorithm with a wide range of kernel types including Gaussian kernels converge if the problem is one dimensional.

In this paper, we provide a theorem for convergence of mode estimate sequences generated by the MS-type algorithms. The proof relies on analyticity of the kernel function, while not requiring assumptions either on the finiteness of stationary points of an estimated PDF, on non-degeneracy of the Hessian of an estimate PDF at stationary points, or on sufficiently small step sizes. This theorem therefore covers most of typical settings appearing in practice, including the case where the Gaussian kernel is used in estimating the PDF.

Also, despite many studies on the convergence of the MS algorithm, how the mode/density estimate sequences behave have not been clarified yet. In particular, in view of the MS algorithm as a gradient ascent method, significance of the particular choice of step sizes adopted in the conventional MS algorithm has not yet been fully elucidated. In this paper, we study the MS-type algorithms with an additional parameter to control step sizes, and show that density estimates along the sequence of mode estimates obtained from the MS-type algorithms are non-decreasing if each step size is up to twice as large as that used in the conventional MS-type algorithms. Moreover, the above-mentioned convergence conditions of the mode estimate sequences remain true even if the step sizes up to about twice the conventional step size are used. These results suggest that the computational efficiency of the MS-type algorithms may be improved practically by using step sizes larger than the conventional ones. Another result is that the density estimate monotonically increases on the line segment between two consecutive mode estimates generated by the MS-type algorithms using a Gaussian kernel and the conventional step size. In the one-dimensional case, it implies that the mode estimate sequence monotonically converges. On the basis of these results, we propose two acceleration techniques of the MS-type algorithms.

The organization of this paper is as follows. First, we explain the MS algorithm and its variants in Section 2. Next, we theoretically analyze these algorithms in Section 3. Then, in Section 4, we propose two acceleration techniques of these algorithms on the basis of the theoretical results in Section 3 and confirm their improvement via numerical experiments. Finally, the conclusions are given in Section 5.

2 Mean Shift Algorithm and its Variants

2.1 Mean Shift Algorithm

Let $\boldsymbol{X} \in \mathbb{R}^d$ be a random variable. Assume that the probability distribution of $\boldsymbol{X}$ has a PDF p. Modes of a PDF p are defined as local maximizers of p. In practical situations, one often cannot have direct access to the PDF p itself but only a finite number of samples drawn from it are available. In such cases, although it is not possible to know exactly the PDF p and its modes, one can still consider estimating the PDF on the basis of the samples available, and then estimating the modes as local maximizers of an estimated PDF.

For estimating the PDF, the MS algorithm typically uses kernel density estimation, which is one of the most representative nonparametric density estimation methods. Assume that a sample set $\mathcal{D} \coloneqq \{\boldsymbol{x}_i \in \mathbb{R}^d\}_{i=1}^n$, consisting of a finite number n of independent and identically-distributed samples drawn from the probability distribution with the PDF p, is available. We define a weight set as $\mathcal{W} \coloneqq \{w_i > 0\}_{i=1}^n$. The kernel density estimate (KDE) of the PDF p is then given by

$$\hat{p}_{\mathcal{D},\mathcal{W}}(\boldsymbol{x}) \coloneqq \sum_{i=1}^{n} w_i K(\boldsymbol{x} - \boldsymbol{x}_i), \tag{1}$$

where K is the kernel function and where $w_i > 0$ is the weight for sample i. In the following, we will drop the subscripts $\mathcal{D}$ and $\mathcal{W}$ of $\hat{p}_{\mathcal{D},\mathcal{W}}(\boldsymbol{x})$ when they are obvious from the context. In the following discussion, the kernel function K is assumed differentiable, nonnegative, and normalized, that is, $\int_{\mathbb{R}^d} K(\boldsymbol{x})\, d\boldsymbol{x} = 1$ holds. Note that (1) reduces to the conventional KDE by letting $w_i = 1/n$, $i = 1, \ldots, n$. Note also that in the definition above the KDE is not normalized unless $\sum_{i=1}^n w_i = 1$ holds, so that the following

discussion also covers the generalized density estimate and the associated weighted MS discussed in [28].

In kernel density estimation, a scale parameter (bandwidth) of the kernel function significantly affects accuracy of the density estimate. It is therefore important to select an appropriate bandwidth in kernel density estimation as well as in the MS-type algorithms [29], [30]. In this paper, however, we assume that the bandwidth has already been appropriately determined, so that we do not discuss how to determine it. Also, although one does not have to use the same bandwidth for the n appearances of the kernel function in (1), we assume, unless otherwise stated, that the same bandwidth is used for all appearances of the kernel function.

One observes that kernel density estimation is translation invariant in the following sense. Given a sample set $\mathcal{D}$, a weight set $\mathcal{W}$, and an arbitrary constant vector $\boldsymbol{a} \in \mathbb{R}^d$, consider the translated sample set $\mathcal{D}_{\boldsymbol{a}} = \{\boldsymbol{x}_i + \boldsymbol{a}\}_{i=1}^n$. One then has

$$\hat{p}_{\mathcal{D}_{\boldsymbol{a}},\mathcal{W}}(\boldsymbol{x} + \boldsymbol{a}) = \hat{p}_{\mathcal{D},\mathcal{W}}(\boldsymbol{x}). \tag{2}$$

According to the usual discussion, assume that the kernel function K is radially symmetric and strictly decreasing with respect to the Euclidean norm of the argument. Then, it can be expressed as $K(\boldsymbol{x}) = k(\|\boldsymbol{x}\|^2/2)$ using a certain function k, called the profile of the kernel function K, where $\|\cdot\|$ denotes the Euclidean norm. Let $g(u) = -k'(u)$, and define a new kernel function G as $G(\boldsymbol{x}) \coloneqq g(\|\boldsymbol{x}\|^2/2)$ with g its profile. For example, if K is a Gaussian kernel, G is also a Gaussian kernel. It should be noted that G might not be normalized even if K is normalized. Since $\nabla K(\boldsymbol{x}) = -\boldsymbol{x}G(\boldsymbol{x})$ holds, the gradient of the KDE $\hat{p}$ is

$$\nabla\hat{p}(\boldsymbol{x}) = \sum_{i=1}^n (\boldsymbol{x}_i - \boldsymbol{x}) w_i G(\boldsymbol{x} - \boldsymbol{x}_i). \tag{3}$$

Given a differentiable PDF estimate $\hat{p}$, a mode estimate should satisfy $\nabla\hat{p}(\boldsymbol{x}) = \boldsymbol{0}$. When $\hat{p}$ is given as a KDE, this condition is rewritten as

$$\sum_{i=1}^n (\boldsymbol{x}_i - \boldsymbol{x}) w_i G(\boldsymbol{x} - \boldsymbol{x}_i) = \boldsymbol{0}. \tag{4}$$

The MS algorithm is derived as a fixed-point iteration on the basis of the above condition, as

$$\boldsymbol{y}_{t+1} = \frac{\sum_{i=1}^n \boldsymbol{x}_i w_i G(\boldsymbol{y}_t - \boldsymbol{x}_i)}{\sum_{i=1}^n w_i G(\boldsymbol{y}_t - \boldsymbol{x}_i)}. \tag{5}$$

We call $\{\boldsymbol{y}_t\}$ a mode estimate sequence with an initial mode estimate $\boldsymbol{y}_0$. We also call the sequence $\{\hat{p}(\boldsymbol{y}_t)\}$ the density estimate sequence.

The iterative equation derived above can alternatively be viewed as a gradient ascent method using an adaptive step size. The gradient ascent method is formulated as

$$\boldsymbol{y}_{t+1} = \boldsymbol{y}_t + \eta_t \nabla\hat{p}(\boldsymbol{y}_t), \tag{6}$$

where $\eta_t > 0$ is the step size at iteration t. If one chooses

$$\eta_t = \frac{1}{\sum_{i=1}^n w_i G(\boldsymbol{y}_t - \boldsymbol{x}_i)} > 0, \tag{7}$$

then the expression (6) reduces to (5), implying that the MS algorithm is a gradient ascent method[1].

For use in the following discussion, we define the MS function as

$$\boldsymbol{m}_{\mathcal{D},\mathcal{W}}(\boldsymbol{x}) \coloneqq \frac{\sum_{i=1}^n \boldsymbol{x}_i w_i G(\boldsymbol{x} - \boldsymbol{x}_i)}{\sum_{i=1}^n w_i G(\boldsymbol{x} - \boldsymbol{x}_i)} - \boldsymbol{x}. \tag{8}$$

In the following, we will drop the subscripts $\mathcal{D}$ and $\mathcal{W}$ of $\boldsymbol{m}_{\mathcal{D},\mathcal{W}}(\boldsymbol{x})$ when they are obvious from the context. The MS algorithm can be reformulated using the MS function as $\boldsymbol{y}_{t+1} = \boldsymbol{y}_t + \boldsymbol{m}(\boldsymbol{y}_t)$. One also has $\boldsymbol{m}(\boldsymbol{x}) = \nabla\hat{p}(\boldsymbol{x})/(\sum_{i=1}^n w_i G(\boldsymbol{x} - \boldsymbol{x}_i))$, that is, the MS function $\boldsymbol{m}(\boldsymbol{x})$ is proportional to the gradient $\nabla\hat{p}(\boldsymbol{x})$. It should be noted that our definition of the MS function given above reduces to the conventional definition, which assumes the uniform weights, that is, $w_i = 1/n$, $i = 1, \ldots, n$.

The translation invariance property of kernel density estimation discussed above carries over to the MS function: Given a sample set $\mathcal{D}$, a weight set $\mathcal{W}$, and an arbitrary constant vector $\boldsymbol{a} \in \mathbb{R}^d$, one has

$$\boldsymbol{m}_{\mathcal{D}_{\boldsymbol{a}},\mathcal{W}}(\boldsymbol{x} + \boldsymbol{a}) = \boldsymbol{m}_{\mathcal{D},\mathcal{W}}(\boldsymbol{x}). \tag{9}$$

If G is positive-valued, for any $\boldsymbol{x}$,

$$\boldsymbol{x} + \boldsymbol{m}_{\mathcal{D},\mathcal{W}}(\boldsymbol{x}) = \frac{\sum_{i=1}^n \boldsymbol{x}_i w_i G(\boldsymbol{x} - \boldsymbol{x}_i)}{\sum_{i=1}^n w_i G(\boldsymbol{x} - \boldsymbol{x}_i)} \tag{10}$$

is a convex combination of the samples in $\mathcal{D}$. One then has:

Lemma 1. *Assume that a kernel function K has a differentiable and strictly decreasing profile. Let $\mathcal{C}$ be the convex hull of the sample set $\mathcal{D}$. One then has $\boldsymbol{x} + \boldsymbol{m}_{\mathcal{D},\mathcal{W}}(\boldsymbol{x}) \in \mathcal{C}$, for any $\boldsymbol{x}$.*

2.2 Variants of the Mean Shift Algorithm

Various variants of the MS algorithm have been proposed. The CMS algorithm can be regarded as a weighted version of the MS algorithm having the weights w_i related to the independent variable part of the sample points $\boldsymbol{x}_i$ (see [11], [12], [13] for details). Therefore, the CMS algorithm is included in the general MS algorithm (5) as a special case.

For the SCMS algorithm, we leave its details to [15], [16], [17]. Here we briefly review the iteration formula of the SCMS algorithm. The SCMS algorithm iterates

$$\boldsymbol{y}_{t+1} = \boldsymbol{y}_t + \mathbf{U}_t \boldsymbol{m}(\boldsymbol{y}_t), \tag{11}$$

where $\mathbf{U}_t$ is typically chosen as the projection matrix representing the normal projection onto the subspace spanned by the $(d - j)$ eigenvectors, associated with the $(d - j)$ largest eigenvalues, of the negative Hessian of $\log\hat{p}$ at $\boldsymbol{y}_t$, which is called the local inverse covariance matrix. When $\mathbf{U}_t$ is an identity matrix, the iteration formula of the SCMS algorithm is identical to those of the MS algorithm and the CMS algorithm, so that the SCMS algorithm can be regarded as a generalization of the MS algorithm. As methods for mode estimation, MS-variants based on Newton's method that uses $\mathbf{U}_t$ proportional to $(\nabla^2\hat{p}(\boldsymbol{y}_t))^{-1}$ have also been developed [31].

1. Historically, the MS algorithm was originally proposed as a gradient ascent method as in (6) [1, Section IV-A]. The particular form (5) of the MS algorithm, or equivalently, the particular choice (7) of the step size, was studied extensively in [2], and it is this particular form that has become commonly used.

3 Properties and Convergence of the Mean Shift-Type Algorithm

3.1 Properties of the Kernel Density Estimate

Before studying properties of the mode/density estimate sequences generated by the MS-type algorithms, we investigate properties of the KDE in this subsection.

First, we show that the KDE has no stationary point outside the convex hull $\mathcal{C}$ of the sample set. This fact, stated in Proposition 1 below, justifies the behavior of the MS algorithm implied by Lemma 1 that it seeks modes only inside $\mathcal{C}$.

Proposition 1. *Assume that a kernel function K has a differentiable and strictly decreasing profile. Then, the gradient $\nabla\hat{p}$ of the KDE is nonzero outside the convex hull $\mathcal{C}$ of the sample set $\mathcal{D}$.*

Note. The Gaussian-kernel case, and the uniform-weight case $w_i = 1/n$, $i = 1, \ldots, n$, of this proposition have been proved in [32] and [27, Lemma 1], respectively. A closely related result is [33, Theorem 2.1], which, without assuming differentiability of the profile, proved absence of a maximum, not of a stationary point, of the KDE outside $\mathcal{C}$.

Proof: Let $\boldsymbol{y} \notin \mathcal{C}$ be an arbitrary point outside $\mathcal{C}$. Since $\mathcal{D}$ is finite, $\mathcal{C}$ is bounded and closed, so that there exists $\boldsymbol{x}_0 \in \mathcal{C}$ which is the closest to $\boldsymbol{y}$. The gradient $\nabla\hat{p}$ of the KDE at $\boldsymbol{y}$ is given by

$$\nabla\hat{p}(\boldsymbol{y}) = -\sum_{i=1}^{n}(\boldsymbol{y} - \boldsymbol{x}_i)w_i G(\boldsymbol{y} - \boldsymbol{x}_i). \tag{12}$$

By assumption, one has $G(\boldsymbol{y} - \boldsymbol{x}_i) > 0$ and $w_i > 0$ for all $i = 1, \ldots, n$. From Lemma 4 in Appendix, $(\boldsymbol{y} - \boldsymbol{x}_0) \cdot (\boldsymbol{y} - \boldsymbol{x}_i) > 0$ holds for all $i = 1, \ldots, n$. Thus, one has proved $(\boldsymbol{y} - \boldsymbol{x}_0) \cdot \nabla\hat{p}(\boldsymbol{y}) < 0$ at any point $\boldsymbol{y} \notin \mathcal{C}$, and hence $\nabla\hat{p}$ does not vanish outside $\mathcal{C}$. □

Also, we show that the KDE using an analytic kernel function does not have a plateau defined below.

Definition 1 (Plateau)**.** *A* plateau *of a function f on $S \subset \mathbb{R}^d$ is defined as an open subset of S where f takes a constant value.*

Proposition 2. *Assume that a kernel function K is analytic. Then, the KDE $\hat{p}$ has no plateau.*

Proof: Assume to the contrary that $\hat{p}$ is constant on an open set $S \subset \mathbb{R}^d$. Take a line segment $\ell = \{\boldsymbol{y}(\epsilon) = \boldsymbol{y}_0 + \epsilon\boldsymbol{d} : \epsilon \in (0, 1)\} \subset S$, $\boldsymbol{d} \neq \boldsymbol{0}$, and define $f : (0, 1) \to \mathbb{R}$ as $f(\epsilon) = \hat{p}(\boldsymbol{y}(\epsilon))$. By the analyticity of K, the function f is also an analytic function. Treating ϵ as a complex-valued variable, the expression of ϵ then defines a function which is holomorphic on $\mathbb{C}$ (i.e., a complex function that is analytic on $\mathbb{C}$) and is equal to $\hat{p}$ on the straight line containing the line segment ℓ. Since f is holomorphic on $\mathbb{C}$ and takes a constant value on the interval $(0, 1)$ in $\mathbb{C}$, from the identity theorem[2], f is constant throughout $\mathbb{C}$. But it is impossible because $f(\epsilon) = \hat{p}(\boldsymbol{y}(\epsilon))$ as a function of real-valued variable ϵ should be nonzero and should decay toward zero as $|\epsilon| \to \infty$, which is contradiction. □

In the one-dimensional case, a similar argument has been adopted for a Gaussian kernel in [16, Proposition 1] to show the finiteness of the stationary points of $\hat{p}$. This argument based on the identity theorem in complex analysis to prove the finiteness of the stationary points does not extend to two or more dimensions, however, since zero sets of analytic functions in more than one variable are known to be nondiscrete. There are studies on properties of the MS algorithm on the basis of theories of a Morse function (i.e., a function having no degenerate stationary points): See, e.g., [34]. Even though the number of non-degenerate stationary points of the KDE with a Gaussian kernel has been shown to be finite [26], whether the total number of stationary points of the KDE with a Gaussian kernel, including degenerate ones, is finite is still an open problem.

2. The identity theorem states that, for two holomorphic functions f and g in a region $\Omega \subset \mathbb{C}$, if $f(z) = g(z)$ holds for all z in a set which has an accumulation point in Ω, then $f(z) = g(z)$ holds for all $z \in \Omega$.

3.2 Properties of the Mean Shift Function

In this subsection, we discuss properties of the MS function, which will be utilized in the following sections to study properties of the MS-type algorithms. Here we introduce novel concepts related to the MS function.

Definition 2 (Improvement ball and ascent ball)**.** *Given the MS function $\boldsymbol{m}$ and a point $\boldsymbol{x} \in \mathbb{R}^d$, we define the* improvement ball *$\mathcal{I}(\boldsymbol{x})$ at $\boldsymbol{x}$ associated with the MS function $\boldsymbol{m}(\boldsymbol{x})$ as the d-dimensional ball centered at $\boldsymbol{x} + \boldsymbol{m}(\boldsymbol{x})$ and of radius $\|\boldsymbol{m}(\boldsymbol{x})\|$. Similarly, we define the* ascent ball *$\mathcal{A}(\boldsymbol{x})$ at $\boldsymbol{x}$ associated with the MS function $\boldsymbol{m}(\boldsymbol{x})$ as the d-dimensional ball centered at $\boldsymbol{x} + \boldsymbol{m}(\boldsymbol{x})/2$ and of radius $\|\boldsymbol{m}(\boldsymbol{x})\|/2$.*

Note that these balls are *not* centered at $\boldsymbol{x}$. Examples of these balls can be found in Fig. 1. The significance of these definitions will be elucidated in the rest of this subsection. Briefly, if the kernel function K satisfies certain properties, then one can guarantee that the value of $\hat{p}$ in the improvement ball $\mathcal{I}(\boldsymbol{x})$ is at least as large as $\hat{p}(\boldsymbol{x})$ (Lemma 2). The improvement ball $\mathcal{I}(\boldsymbol{x})$ is also related to convergence theorems (Theorems 1 and 2), given in the next subsection, of the density/mode estimate sequences. Furthermore, if K is a Gaussian kernel, then one can guarantee that, in the ascent ball $\mathcal{A}(\boldsymbol{x})$, $\hat{p}$ is increasing along any line passing through $\boldsymbol{x}$ (Lemma 3). In particular, $\nabla\hat{p}$ is shown to be non-vanishing in the interior of $\mathcal{A}(\boldsymbol{x})$.

First, we provide a sufficient condition for $\boldsymbol{y}$ given $\boldsymbol{x}$, in terms of the improvement ball $\mathcal{I}(\boldsymbol{x})$ at $\boldsymbol{x}$, such that the inequality $\hat{p}(\boldsymbol{y}) > \hat{p}(\boldsymbol{x})$ holds.

Lemma 2. *Assume that a kernel function K has a differentiable and strictly decreasing profile. For a point $\boldsymbol{x} \in \mathbb{R}^d$, let $\mathcal{I}(\boldsymbol{x})$ be the improvement ball at $\boldsymbol{x}$ associated with the MS function $\boldsymbol{m}(\boldsymbol{x})$. Then, for any $\boldsymbol{y} \in \mathcal{I}(\boldsymbol{x})$, one has $\hat{p}(\boldsymbol{y}) \geq \hat{p}(\boldsymbol{x})$, with strict inequality if $\boldsymbol{y}$ is an interior point of $\mathcal{I}(\boldsymbol{x})$.*

Proof: The translation invariance property of kernel density estimation allows us to assume $\boldsymbol{x} = \boldsymbol{0}$ without loss of generality. The difference of the density estimate at $\boldsymbol{y}$ and that at $\boldsymbol{0}$ is given by

$$\hat{p}(\boldsymbol{y}) - \hat{p}(\boldsymbol{0}) = \sum_{i=1}^{n} w_i \left[k\left(\frac{\|\boldsymbol{y} - \boldsymbol{x}_i\|^2}{2}\right) - k\left(\frac{\|\boldsymbol{x}_i\|^2}{2}\right)\right]. \tag{13}$$

The convexity and the differentiability of the profile k, as well as the definition of g, yields the inequality $k(u_2) \geq$

$k(u_1) + g(u_1)(u_1 - u_2)$ to hold for any $u_1, u_2 \in [0, \infty)$. It allows us to provide a lower bound of $\hat{p}(\boldsymbol{y}) - \hat{p}(\boldsymbol{0})$, as

$$\begin{aligned}
&\hat{p}(\boldsymbol{y}) - \hat{p}(\boldsymbol{0}) \\
&\geq \frac{1}{2}\sum_{i=1}^{n} w_i G(\boldsymbol{x}_i)\left(\|\boldsymbol{x}_i\|^2 - \|\boldsymbol{y} - \boldsymbol{x}_i\|^2\right) \\
&= \boldsymbol{y} \cdot \sum_{i=1}^{n} \boldsymbol{x}_i w_i G(\boldsymbol{x}_i) - \frac{1}{2}\|\boldsymbol{y}\|^2 \sum_{i=1}^{n} w_i G(\boldsymbol{x}_i) \\
&= \left(\boldsymbol{y} \cdot \boldsymbol{m}(\boldsymbol{0}) - \frac{1}{2}\|\boldsymbol{y}\|^2\right) \sum_{i=1}^{n} w_i G(\boldsymbol{x}_i) \\
&= \frac{1}{2}\left(\|\boldsymbol{m}(\boldsymbol{0})\|^2 - \|\boldsymbol{y} - \boldsymbol{m}(\boldsymbol{0})\|^2\right) \sum_{i=1}^{n} w_i G(\boldsymbol{x}_i). \qquad (14)
\end{aligned}$$

Since the improvement ball $\mathcal{I}(\boldsymbol{0})$ is centered at $\boldsymbol{m}(\boldsymbol{0})$ and of radius $\|\boldsymbol{m}(\boldsymbol{0})\|$, one has $\|\boldsymbol{y} - \boldsymbol{m}(\boldsymbol{0})\|^2 \leq \|\boldsymbol{m}(\boldsymbol{0})\|^2$ for any $\boldsymbol{y} \in \mathcal{I}(\boldsymbol{0})$. Also, since the profile k is assumed strictly decreasing, $g(u) = -k'(u)$ is positive, and so is the kernel G. Consequently, the last line of (14) is nonnegative for any $\boldsymbol{y} \in \mathcal{I}(\boldsymbol{0})$, and is strictly positive if $\boldsymbol{y}$ is an interior point of $\mathcal{I}(\boldsymbol{0})$. One has therefore shown that $\hat{p}(\boldsymbol{y}) \geq \hat{p}(\boldsymbol{x})$ holds for any $\boldsymbol{y} \in \mathcal{I}(\boldsymbol{x})$, with strict inequality when $\boldsymbol{y}$ is an interior point of $\mathcal{I}(\boldsymbol{x})$. □

Next, we provide, under the condition that a Gaussian kernel is used, a sufficient condition for $\boldsymbol{y}$ given $\boldsymbol{x}$, in terms of the ascent ball $\mathcal{A}(\boldsymbol{x})$ at $\boldsymbol{x}$, such that $\nabla\hat{p}(\boldsymbol{y})$ is non-vanishing.

Lemma 3. *Assume that a kernel function K is a Gaussian kernel. For a point $\boldsymbol{x} \in \mathbb{R}^d$, let $\mathcal{A}(\boldsymbol{x})$ be the ascent ball at $\boldsymbol{x}$ associated with the MS function $\boldsymbol{m}(\boldsymbol{x})$. Then, for any $\boldsymbol{y} \in \mathcal{A}(\boldsymbol{x})$, the inner product of the gradient $\nabla\hat{p}(\boldsymbol{y})$ of the KDE at $\boldsymbol{y}$ and $(\boldsymbol{y} - \boldsymbol{x})$ is nonnegative. Moreover, if $\boldsymbol{y}$ is an interior point of $\mathcal{A}(\boldsymbol{x})$, or if there is at least one sample $\boldsymbol{x}_i$ such that $(\boldsymbol{y} - \boldsymbol{x}) \cdot (\boldsymbol{y} - \boldsymbol{x}_i)$ is nonzero, the above inner product is strictly positive.*

Proof: Let

$$f(\boldsymbol{y}) := \nabla\hat{p}(\boldsymbol{y}) \cdot (\boldsymbol{y} - \boldsymbol{x}). \qquad (15)$$

We wish to prove the nonnegativity of $f(\boldsymbol{y})$. For $\boldsymbol{y} = \boldsymbol{x}$, it is trivially zero. For $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{m}(\boldsymbol{x})$, the positivity of f has been proved in [3, Theorem 2]. Now the question is whether $f(\boldsymbol{y})$ is nonnegative for other values of $\boldsymbol{y} \in \mathcal{A}(\boldsymbol{x})$.

We again make use of the translation invariance to assume $\boldsymbol{x} = \boldsymbol{0}$ without loss of generality. One has

$$f(\boldsymbol{y}) = \sum_{i=1}^{n} \boldsymbol{y} \cdot (\boldsymbol{x}_i - \boldsymbol{y}) w_i G(\boldsymbol{y} - \boldsymbol{x}_i), \qquad (16)$$

of which we wish to prove the nonnegativity.

We start with the definition of the MS function $\boldsymbol{m}(\boldsymbol{x})$, which is rewritten when $\boldsymbol{x} = \boldsymbol{0}$ as

$$\sum_{i=1}^{n} (\boldsymbol{m}(\boldsymbol{0}) - \boldsymbol{x}_i) w_i G(\boldsymbol{x}_i) = \boldsymbol{0}. \qquad (17)$$

Now, consider the quantity

$$\sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{y}) w_i G(\boldsymbol{y} - \boldsymbol{x}_i). \qquad (18)$$

Adding to it the left-hand side of (17) multiplied by $G(\boldsymbol{0})/G(\boldsymbol{y})$ does not change its value, so that one has

$$\sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{y}) w_i G(\boldsymbol{y} - \boldsymbol{x}_i) = \frac{G(\boldsymbol{0})}{G(\boldsymbol{y})} \sum_{i=1}^{n} A(\boldsymbol{x}_i, \boldsymbol{y}) w_i G(\boldsymbol{x}_i), \qquad (19)$$

where we let

$$\begin{aligned}
A(\boldsymbol{x}, \boldsymbol{y}) &= (\boldsymbol{x} - \boldsymbol{y}) \frac{G(\boldsymbol{y} - \boldsymbol{x}) G(\boldsymbol{y})}{G(\boldsymbol{x}) G(\boldsymbol{0})} + (\boldsymbol{m}(\boldsymbol{0}) - \boldsymbol{x}) \\
&= (\boldsymbol{x} - \boldsymbol{y}) \left[\frac{G(\boldsymbol{y} - \boldsymbol{x}) G(\boldsymbol{y})}{G(\boldsymbol{x}) G(\boldsymbol{0})} - 1\right] + (\boldsymbol{m}(\boldsymbol{0}) - \boldsymbol{y}). \qquad (20)
\end{aligned}$$

Since G is a Gaussian kernel, one can write it as $G(\boldsymbol{x}) = a \exp(-b\|\boldsymbol{x}\|^2/2)$ with $a, b > 0$. One then has

$$\frac{G(\boldsymbol{y} - \boldsymbol{x}) G(\boldsymbol{y})}{G(\boldsymbol{x}) G(\boldsymbol{0})} = \exp[-b\boldsymbol{y} \cdot (\boldsymbol{y} - \boldsymbol{x})]. \qquad (21)$$

Taking the inner product of $A(\boldsymbol{x}, \boldsymbol{y})$ and $\boldsymbol{y}$ gives

$$\begin{aligned}
\boldsymbol{y} \cdot A(\boldsymbol{x}, \boldsymbol{y}) = &- \boldsymbol{y} \cdot (\boldsymbol{y} - \boldsymbol{x})[\exp(-b\boldsymbol{y} \cdot (\boldsymbol{y} - \boldsymbol{x})) - 1] \\
&- \boldsymbol{y} \cdot (\boldsymbol{y} - \boldsymbol{m}(\boldsymbol{0})). \qquad (22)
\end{aligned}$$

Since $t(e^{bt} - 1) \geq 0$ for all $t \in \mathbb{R}$, one has

$$-\boldsymbol{y} \cdot (\boldsymbol{y} - \boldsymbol{x})[\exp(-b\boldsymbol{y} \cdot (\boldsymbol{y} - \boldsymbol{x})) - 1] \geq 0. \qquad (23)$$

When $\boldsymbol{y} \in \mathcal{A}(\boldsymbol{0})$, one has $\boldsymbol{y} \cdot (\boldsymbol{y} - \boldsymbol{m}(\boldsymbol{0})) = \|\boldsymbol{y} - \boldsymbol{m}(\boldsymbol{0})/2\|^2 - \|\boldsymbol{m}(\boldsymbol{0})/2\|^2 \leq 0$, with strict inequality if and only if $\boldsymbol{y}$ is an interior point of $\mathcal{A}(\boldsymbol{0})$. We have therefore shown that $\boldsymbol{y} \cdot A(\boldsymbol{x}, \boldsymbol{y}) \geq 0$ holds for any $\boldsymbol{x}$, with strict inequality either if $\boldsymbol{y}$ is an interior point of $\mathcal{A}(\boldsymbol{0})$, or if $\boldsymbol{y}$ and $(\boldsymbol{y} - \boldsymbol{x})$ are not orthogonal.

The argument so far, along with the positivity of the kernel G and the weights w_i, has proved that $f(\boldsymbol{y})$, which equals to the inner product of $\boldsymbol{y}$ and the right-hand side of (19), is nonnegative. It is strictly positive either if $\boldsymbol{y}$ is an interior point of $\mathcal{A}(\boldsymbol{0})$, or if there is at least one sample $\boldsymbol{x}_i$ such that $\boldsymbol{y} \cdot (\boldsymbol{y} - \boldsymbol{x}_i)$ is not equal to zero. □

When $\boldsymbol{y} \neq \boldsymbol{x}$, a simple sufficient condition for the existence of a sample $\boldsymbol{x}_i$ satisfying $(\boldsymbol{y} - \boldsymbol{x}) \cdot (\boldsymbol{y} - \boldsymbol{x}_i) \neq 0$ is that the affine space containing the sample set $\mathcal{D}$ with the minimum dimension is $\mathbb{R}^d$.

The arguments on the basis of the improvement ball and the ascent ball described so far tell us how the KDE $\hat{p}$ and its gradient $\nabla\hat{p}$ behave in the vicinity of $\boldsymbol{x}$, but information they provide is "one-sided" in the sense that they tell us nothing about behaviors of $\hat{p}$ or $\nabla\hat{p}$ at $\boldsymbol{y}$ satisfying $(\boldsymbol{y} - \boldsymbol{x}) \cdot \boldsymbol{m}(\boldsymbol{x}) < 0$. "Two-sided" information in this sense may be obtained via an argument on the basis of Lipschitz continuity of gradient of the kernel function. Such information, however, is weak compared with that obtained via the MS function, as discussed below.

Definition 3. *A differentiable function f defined on $S \subset \mathbb{R}^d$ has a Lipschitz-continuous gradient if there exists $L > 0$ such that*

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\| \qquad (24)$$

for any $\boldsymbol{x}, \boldsymbol{y} \in S$, where L is called the Lipschitz constant of ∇f.

Note that if f is of class C^2 then one can take $L = \sup_{\boldsymbol{x} \in S} \|\nabla^2 f(\boldsymbol{x})\|_{\mathrm{op}}$, where $\|\cdot\|_{\mathrm{op}}$ denotes the operator norm induced by the Euclidean norm in $\mathbb{R}^d$.

Definition 4 (Lipschitz ball). *Given a KDE $\hat{p}$ based on a kernel function K, which has a Lipschitz-continuous gradient with Lipschitz constant $L > 0$, and a point $\boldsymbol{x} \in \mathbb{R}^d$, we define the* Lipschitz ball *$\mathcal{L}(\boldsymbol{x})$ at $\boldsymbol{x}$ as the d-dimensional ball centered at $\boldsymbol{x}$ and with radius $\|\nabla\hat{p}(\boldsymbol{x})\|/\left(L\sum_{i=1}^n w_i\right)$.*

Proposition 3. *Assume that a kernel function K has a Lipschitz-continuous gradient with Lipschitz constant L. Assume that the gradient of the KDE at $\boldsymbol{x}$ is nonzero. Then the gradient $\nabla\hat{p}$ is nonzero in the interior of the Lipschitz ball $\mathcal{L}(\boldsymbol{x})$ at $\boldsymbol{x}$.*

Moreover, if K is a Gaussian kernel, the radius of the Lipschitz ball $\mathcal{L}(\boldsymbol{x})$ is strictly less than $\|\boldsymbol{m}(\boldsymbol{x})\|$.

Proof: The gradient $\nabla\hat{p}$ of the KDE is written as

$$\nabla\hat{p}(\boldsymbol{x}) = \sum_{i=1}^n w_i \nabla K(\boldsymbol{x} - \boldsymbol{x}_i). \tag{25}$$

Since K has a Lipschitz-continuous gradient, one has for any $\boldsymbol{x}$ and $\boldsymbol{y}$,

$$\begin{aligned}\|\nabla K(\boldsymbol{x} - \boldsymbol{x}_i) - \nabla K(\boldsymbol{y} - \boldsymbol{x}_i)\| &\le L\|(\boldsymbol{x} - \boldsymbol{x}_i) - (\boldsymbol{y} - \boldsymbol{x}_i)\| \\ &= L\|\boldsymbol{x} - \boldsymbol{y}\|. \end{aligned}\tag{26}$$

Let $\boldsymbol{y}$ be any stationary point of $\hat{p}$, so that $\nabla\hat{p}(\boldsymbol{y}) = \boldsymbol{0}$ holds. One then has

$$\begin{aligned}\|\nabla\hat{p}(\boldsymbol{x})\| &= \|\nabla\hat{p}(\boldsymbol{x}) - \nabla\hat{p}(\boldsymbol{y})\| \\ &\le \sum_{i=1}^n w_i \|\nabla K(\boldsymbol{x} - \boldsymbol{x}_i) - \nabla K(\boldsymbol{y} - \boldsymbol{x}_i)\| \\ &\le L\|\boldsymbol{x} - \boldsymbol{y}\| \sum_{i=1}^n w_i, \end{aligned}\tag{27}$$

where the first inequality is due to the triangle inequality, and where the second inequality follows from (26). We have therefore proved

$$\|\boldsymbol{x} - \boldsymbol{y}\| \ge \frac{\|\nabla\hat{p}(\boldsymbol{x})\|}{L\sum_{i=1}^n w_i}, \tag{28}$$

implying that any stationary point $\boldsymbol{y}$ is not in the interior of the Lipschitz ball $\mathcal{L}(\boldsymbol{x})$. This proves the first part of the proposition.

To prove the remaining part of the proposition, since K is assumed to be a Gaussian kernel, one can let

$$K(\boldsymbol{x}) = a\exp\left(-\frac{b\|\boldsymbol{x}\|^2}{2}\right), \quad b > 0, \quad a = \left(\frac{b}{2\pi}\right)^{d/2}. \tag{29}$$

The gradient and the Hessian of K are calculated as

$$\nabla K(\boldsymbol{x}) = -b\boldsymbol{x}K(\boldsymbol{x}) = -\boldsymbol{x}G(\boldsymbol{x}), \tag{30}$$

$$\nabla^2 K(\boldsymbol{x}) = b(b\boldsymbol{x}\boldsymbol{x}^T - \mathbf{I})K(\boldsymbol{x}), \tag{31}$$

where $\mathbf{I}$ denotes the identity matrix. The Hessian $\nabla^2 K(\boldsymbol{x})$ has eigenvalues $b(b\|\boldsymbol{x}\|^2 - 1)K(\boldsymbol{x})$ (with eigenvector $\boldsymbol{x}$) and $-bK(\boldsymbol{x})$ (with eigenvectors orthogonal to $\boldsymbol{x}$), with the latter being $(d-1)$-fold degenerate. Since the Hessian $\nabla^2 K(\boldsymbol{x})$ is symmetric, its operator norm is equal to the largest absolute value of the eigenvalues, and is therefore given by $bK(\boldsymbol{x})$ if $\|\boldsymbol{x}\|^2 < 2/b$, and $b(b\|\boldsymbol{x}\|^2 - 1)K(\boldsymbol{x})$ if $\|\boldsymbol{x}\|^2 \ge 2/b$. It depends on $\boldsymbol{x}$ but is bounded from above by $L := ab$, since one has $K(\boldsymbol{x}) \le K(\boldsymbol{0}) = a$ and $(b\|\boldsymbol{x}\|^2 - 1)K(\boldsymbol{x}) \le (b\|\boldsymbol{x}\|^2 - 1)K(\boldsymbol{x})|_{\|\boldsymbol{x}\|^2 = 3/b} = 2ae^{-3/2} \approx 0.446a < a$.

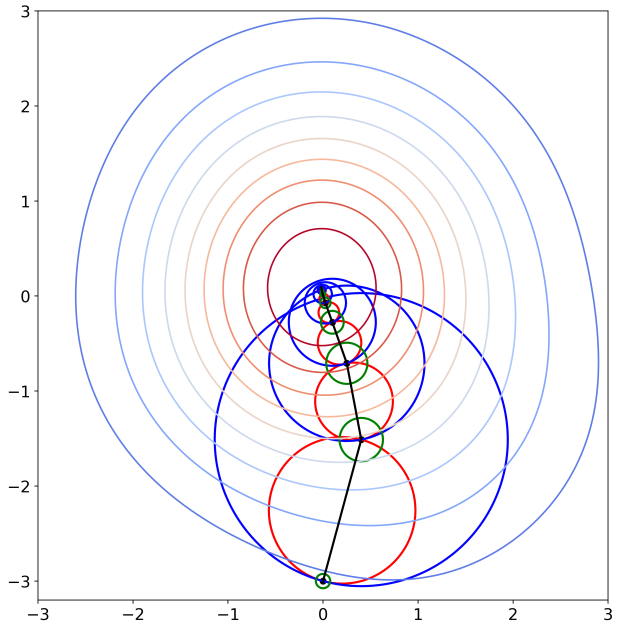

Fig. 1. Comparison of improvement/ascent/Lipschitz balls for the experiment in Example 1. The KDE is plotted as contour lines: the KDE is higher for red and is lower for blue. The black points represent mode estimate sequence $\{\boldsymbol{y}_t\}_{t=0,\dots,10}$. The blue/red/green circles represent the improvement/ascent/Lipschitz balls, respectively.

Here we are interested in relationship between $\|\nabla\hat{p}(\boldsymbol{x})\|/\left(L\sum_{i=1}^n w_i\right)$ and $\|\boldsymbol{m}(\boldsymbol{x})\|$. Since $G(\boldsymbol{x}) = bK(\boldsymbol{x})$ holds, one has

$$\boldsymbol{m}(\boldsymbol{x}) = \frac{\sum_{i=1}^n (\boldsymbol{x}_i - \boldsymbol{x}) w_i K(\boldsymbol{x} - \boldsymbol{x}_i)}{\sum_{i=1}^n w_i K(\boldsymbol{x} - \boldsymbol{x}_i)} = \frac{\nabla\hat{p}(\boldsymbol{x})}{b\hat{p}(\boldsymbol{x})}. \tag{32}$$

Since not all $\|\boldsymbol{x} - \boldsymbol{x}_i\|^2$, $i = 1, \dots, n$, are simultaneously zero when the gradient of $\hat{p}(\boldsymbol{x})$ is nonzero, one has

$$\begin{aligned} b\hat{p}(\boldsymbol{x}) &= b\sum_{i=1}^n w_i a \exp\left(-\frac{b\|\boldsymbol{x} - \boldsymbol{x}_i\|^2}{2}\right) \\ &= L\sum_{i=1}^n w_i \exp\left(-\frac{b\|\boldsymbol{x} - \boldsymbol{x}_i\|^2}{2}\right) < L\sum_{i=1}^n w_i, \end{aligned}\tag{33}$$

which, together with (32), implies that $\|\nabla\hat{p}(\boldsymbol{x})\|/\left(L\sum_{i=1}^n w_i\right)$ is strictly less than $\|\boldsymbol{m}(\boldsymbol{x})\|$, completing the proof for the latter half of the proposition. □

Remark 1. The relationship between the size of the ascent ball and that of the Lipschitz ball depends on the inequality (33). In situations where samples are widely dispersed, this inequality is not tight, so that the radius of the Lipschitz ball can be considerably smaller than the diameter of the ascent ball, as can be observed in Example 1 below.

Example 1. A sample set $\mathcal{D}$ of size 50 was generated from the normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$ on $\mathbb{R}^2$ with mean $\mathbf{0}$ and covariance $\mathbf{I}$. We used the uniform weights $w_i = 1/50$, $i = 1, \dots, 50$, and the Gaussian kernel $K(\boldsymbol{x}) = (2\pi)^{-1}e^{-\|\boldsymbol{x}\|^2/2}$ in kernel density estimation. We iterated $\boldsymbol{y}_{t+1} = \boldsymbol{y}_t + \boldsymbol{m}(\boldsymbol{y}_t)$ ten times from the initial mode estimate $\boldsymbol{y}_0 = (0, -3)^T$ to obtain a mode estimate sequence $\{\boldsymbol{y}_t\}_{t=0,\dots,10}$.

Figure 1 shows the KDE $\hat{p}$, the mode estimate sequence $\{\boldsymbol{y}_t\}_{t=0,\dots,10}$, and the improvement/ascent/Lipschitz balls

at each of $\{\boldsymbol{y}_t\}_{t=0,\dots,10}$. It can be seen that the radius of the Lipschitz ball can be significantly smaller than the diameter of the ascent ball, even in this simple example. □

3.3 Convergence of the Mean Shift-Type Algorithms

In this subsection, we discuss convergence of mode estimate sequences generated by the MS-type algorithms and the corresponding density estimate sequences. For this purpose, we introduce the notions of the mode estimate path and the density estimate path as extensions of the mode estimate sequence and the density estimate sequence, respectively. Properties of mode estimate paths generated by the MS algorithm with sufficiently small step sizes have been studied extensively in [25]. In this paper, these notions will be used in Examples 2, 3, 5, and 6, as well as in Theorem 3.

Definition 5 (Mode estimate path and density estimate path)**.** *Let $\{\boldsymbol{y}_t\}$ be a mode estimate sequence. For $\tau \geq 0$, let $\epsilon = \tau - \lfloor \tau \rfloor$. The* mode estimate path $\boldsymbol{M}(\tau)$ *is defined as*

$$\boldsymbol{M}(\tau) \coloneqq (1-\epsilon)\boldsymbol{y}_{\lfloor \tau \rfloor} + \epsilon \boldsymbol{y}_{\lfloor \tau \rfloor + 1}. \tag{34}$$

The density estimate path $D(\tau)$ *is defined as*

$$D(\tau) \coloneqq \hat{p}(\boldsymbol{M}(\tau)). \tag{35}$$

In other words, the mode estimate path $\boldsymbol{M}(\tau)$ is the piecewise-linear trajectory in $\mathbb{R}^d$ connecting the mode estimates consecutively. The density estimate path $D(\tau)$ is the estimated density values along the mode estimate path. Note that, while $\boldsymbol{M}(\tau)$ is piecewise-linear, $D(\tau)$ is not in general (examples of which will be found in Figs. 2–5).

First, we investigate whether the density estimate sequence, obtained via an extended MS-type algorithm with rescaled step sizes, is non-decreasing.

Theorem 1. *Assume that a kernel function K has a convex, differentiable, and strictly decreasing profile. Let the mode estimate sequence $\{\boldsymbol{y}_t\}$ be obtained via the iteration $\boldsymbol{y}_{t+1} = \boldsymbol{y}_t + \epsilon_t \mathbf{U}_t \boldsymbol{m}(\boldsymbol{y}_t)$, $\epsilon_t \in (0, 2]$, where $\mathbf{U}_t$ is a symmetric projection operator. Then, the density estimate sequence $\{\hat{p}(\boldsymbol{y}_t)\}$ is non-decreasing.*

In addition to the above assumptions, assume further that K is bounded. Then, the sequence $\{\hat{p}(\boldsymbol{y}_t)\}$ converges.

Note. In the case where $\epsilon_t = 1$ for all t, this iteration corresponds to a fixed point iteration such as the conventional MS/CMS/SCMS algorithms, and Theorem 1 in this case has been proved in [16, Proposition 1], [22, Theorem 1].

Proof: If $\mathbf{U}_t \boldsymbol{m}(\boldsymbol{y}_t) = \mathbf{0}$, then one has $\boldsymbol{y}_{t+1} = \boldsymbol{y}_y$ and thus $\hat{p}(\boldsymbol{y}_{t+1}) = \hat{p}(\boldsymbol{y}_t)$. In view of Lemma 2, one has only to show that whenever $\mathbf{U}_t \boldsymbol{m}(\boldsymbol{y}_t) \neq \mathbf{0}$ holds, $\boldsymbol{y}_{t+1}$ is included in the improvement ball $\mathcal{I}(\boldsymbol{y}_t)$ at $\boldsymbol{y}_t$. One has

$$\begin{aligned} &\|\boldsymbol{y}_{t+1} - \boldsymbol{y}_t - \boldsymbol{m}(\boldsymbol{y}_t)\|^2 = \|\epsilon_t \mathbf{U}_t \boldsymbol{m}(\boldsymbol{y}_t) - \boldsymbol{m}(\boldsymbol{y}_t)\|^2 \\ &= (\epsilon_t^2 - 2\epsilon_t)\|\mathbf{U}_t \boldsymbol{m}(\boldsymbol{y}_t)\|^2 + \|\boldsymbol{m}(\boldsymbol{y}_t)\|^2 \leq \|\boldsymbol{m}(\boldsymbol{y}_t)\|^2 \end{aligned} \tag{36}$$

for $\epsilon_t \in (0, 2]$, which implies that $\boldsymbol{y}_{t+1}$ is in $\mathcal{I}(\boldsymbol{y}_t)$. From Lemma 2, one has therefore shown that $\hat{p}(\boldsymbol{y}_{t+1}) \geq \hat{p}(\boldsymbol{y}_t)$ holds for $\epsilon_t \in (0, 2]$.

Also, since the density estimate $\hat{p}$ is bounded under the additional assumption, the sequence $\{\hat{p}(\boldsymbol{y}_t)\}$ converges. □

Remark 2. Fix $\mathbf{U}_t$ and assume $\mathbf{U}_t \boldsymbol{m}(\boldsymbol{y}_t) \neq \mathbf{0}$ to hold. Let ℓ_t be the line passing through $\boldsymbol{y}_t$ and $\boldsymbol{y}_t + \mathbf{U}_t \boldsymbol{m}(\boldsymbol{y}_t)$. Then, the mode estimate $\boldsymbol{y}_{t+1} = \boldsymbol{y}_t + \epsilon_t \mathbf{U}_t \boldsymbol{m}(\boldsymbol{y}_t)$ with $0 < \epsilon_t \leq 2$ satisfies $\boldsymbol{y}_{t+1} \in \ell_t \cap \mathcal{I}(\boldsymbol{y}_t)$. One can show that $\epsilon_t = 1$ maximizes the lower bound of $\hat{p}(\boldsymbol{y}_{t+1}) - \hat{p}(\boldsymbol{y}_t)$ given in (14). Indeed, under the conditions of Theorem 1, $\boldsymbol{y}_t + \mathbf{U}_t \boldsymbol{m}(\boldsymbol{y}_t)$ is the orthogonal projection of $\boldsymbol{y}_t + \boldsymbol{m}(\boldsymbol{y}_t)$ to the line ℓ_t, and thus it minimizes the distance from the center $\boldsymbol{y}_t + \boldsymbol{m}(\boldsymbol{y}_t)$ of the improvement ball $\mathcal{I}(\boldsymbol{y}_t)$ to the line ℓ_t, implying that $\boldsymbol{y}_{t+1} = \boldsymbol{y}_t + \mathbf{U}_t \boldsymbol{m}(\boldsymbol{y}_t)$ maximizes the lower bound (14) of $\hat{p}(\boldsymbol{y}_{t+1}) - \hat{p}(\boldsymbol{y}_t)$ among those $\boldsymbol{y}_{t+1}$s given in the form $\boldsymbol{y}_{t+1} = \boldsymbol{y}_t + \epsilon_t \mathbf{U}_t \boldsymbol{m}(\boldsymbol{y}_t)$ with $0 < \epsilon_t \leq 2$. This result is derived from the convexity of the profile of the kernel function used for the KDE and suggests the adequacy of the adaptive step sizes (i.e., $\epsilon_t = 1$) used in the conventional MS-type algorithms, as experimented in Example 2 below.

Example 2. Let the sample set $\mathcal{D} = \{-0.5, 0.5\}$, the weight set $\mathcal{W} = \{1/3, 2/3\}$, and the initial mode estimate $y_0 = -1.5$. We estimated modes of the KDE $\hat{p}$ using the Gaussian kernel $K(x) = (0.2\pi)^{-1/2} e^{-x^2/0.2}$, which has a bounded, convex, differentiable, and strictly decreasing profile, so that Theorem 1 is applicable. We considered the extended MS algorithm, iterating $y_{t+1} = y_t + \epsilon m(y_t)$ for 10 times. We examined the cases with $\epsilon = 0.5, 1, 1.1, 1.9$, and 2.1.

The KDE $\hat{p}$, the mode estimate path (34), and the density estimate path (35) associated with $\{y_t\}_{t=0,\dots,10}$ are shown in Fig. 2. When $\epsilon > 1$, the mode estimate sequence may not be monotonic. It is confirmed that the corresponding density estimate sequence is monotonically increasing when $\epsilon < 2$, and it may decrease otherwise. □

Remark 3. Example 2 confirms the theoretically-predicted behaviors of the MS algorithm with values of ϵ_t around $\epsilon_t = 1$ and 2, while the theoretical results presented so far do not imply that using $\epsilon_t = 1$ leads to the fastest convergence of the algorithm. For more realistic circumstances where the sample distribution is more dispersed, the inequality (14) may become looser, and then the update by the conventional MS algorithm (i.e., $\epsilon_t = 1$) may tend to be conservative. In such cases, using values of ϵ_t with $\epsilon_t > 1$ may result in faster convergence of the algorithm, as experimented in Example 3 below.

Example 3. Let the sample set $\mathcal{D}$ be of size $n = 5000$ and generated from $(1/3)\mathcal{N}(-0.5, 0.1) + (2/3)\mathcal{N}(0.5, 0.1)$, whose PDF is the same as the KDE in Example 2. We estimated modes of the KDE $\hat{p}$ using the weight $w_i = 1/n$, $i = 1, \dots, n$, and the Gaussian kernel $K(x) = (0.02\pi)^{-1/2} e^{-x^2/0.02}$. We compared the extended algorithms $y_{t+1} = y_t + \epsilon m(y_t)$ with $\epsilon = 1$ and 1.9, starting from $y_0 = -1.5$.

As shown in Fig. 3, the density/mode estimate sequences converged faster when $\epsilon = 1.9$ than when $\epsilon = 1$. □

Corollary 1. *Under the assumptions of Theorem 1, assume further that $\epsilon_t \in [\gamma, 2 - \gamma]$, and that, for any t such that $\boldsymbol{m}(\boldsymbol{y}_t) \neq \mathbf{0}$, $\mathbf{U}_t$ satisfies $\mathbf{U}_t \boldsymbol{m}(\boldsymbol{y}_t) \neq \mathbf{0}$ and*

$$\frac{\mathbf{U}_t \boldsymbol{m}(\boldsymbol{y}_t) \cdot \boldsymbol{m}(\boldsymbol{y}_t)}{\|\mathbf{U}_t \boldsymbol{m}(\boldsymbol{y}_t)\| \|\boldsymbol{m}(\boldsymbol{y}_t)\|} \geq \delta, \tag{37}$$

with t-independent constants $\delta, \gamma \in (0, 1]$. Then, one has

$$\lim_{t \to \infty} \boldsymbol{m}(\boldsymbol{y}_t) = \mathbf{0}. \tag{38}$$

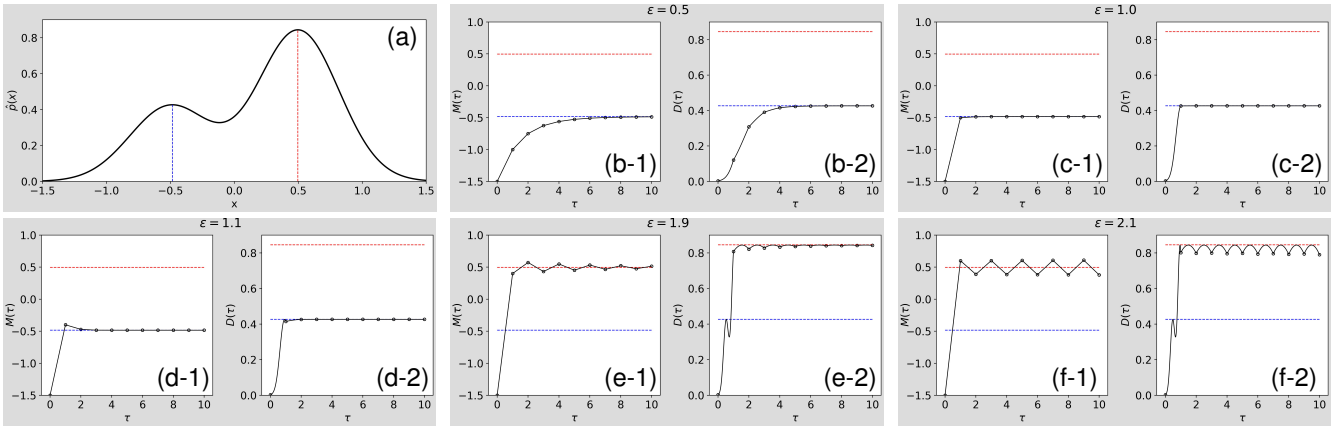

Fig. 2. KDE and mode/density estimate paths for the experiment in Example 2. The red and blue dotted lines represent the locations of the modes near $x = 0.5$ and -0.5, respectively, and the density estimates at these modes. (a) shows the KDE. (b-1),. . .,(f-1) show the mode estimate path $M(\tau)$ in a solid line and the mode estimate sequence $\{y_t\}_{t=0,\ldots,10}$ as points. (b-2),. . .,(f-2) show the density estimate path $D(\tau)$ in a solid line and the density estimate sequence $\{\hat{p}(y_t)\}_{t=0,\ldots,10}$ as points. {(b-1),(b-2)},. . .,{(f-1),(f-2)} are the results with $\epsilon = 0.5, 1, 1.1, 1.9, 2.1$, respectively.

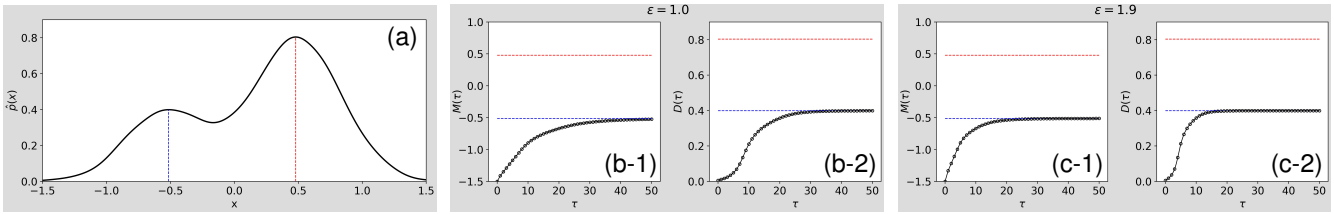

Fig. 3. KDE and mode/density estimate paths for the experiment in Example 3. The red and blue dotted lines represent the locations of the modes near $x = 0.5$ and -0.5, respectively, and the density estimates at these modes. (a) shows the KDE. (b-1) and (c-1) show the mode estimate paths $M(\tau)$ in solid lines and the mode estimate sequences $\{y_t\}_{t=0,\ldots,50}$ as points. (b-2) and (c-2) show the density estimate paths $D(\tau)$ in solid lines and the density estimate sequences $\{\hat{p}(y_t)\}_{t=0,\ldots,50}$ as points. {(b-1),(b-2)} and {(c-1),(c-2)} are the results with $\epsilon = 1, 1.9$, respectively.

If furthermore G is bounded, then

$$\lim_{t\to\infty} \nabla \hat{p}(\boldsymbol{y}_t) = \boldsymbol{0}. \tag{39}$$

Proof: Let $D = \max_{\boldsymbol{x},\boldsymbol{y}\in\mathcal{C}} \|\boldsymbol{x} - \boldsymbol{y}\|$ be the diameter of the convex hull $\mathcal{C}$ of $\mathcal{D}$, let $\Pi_{\mathcal{C}}$ denote the projection operator from $\mathbb{R}^d$ onto $\mathcal{C}$, and for $h \geq 0$, let

$$S_h = \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x} - \Pi_{\mathcal{C}}\boldsymbol{x}\| \leq h\} \tag{40}$$

be the h-neighborhood of $\mathcal{C}$. Also, let us define $\{h_t\}$ via $h_{t+1} = \alpha(h_t + D)$ with $\alpha = \sqrt{1 - \delta^2\gamma(2-\gamma)} \in [0, 1)$ and $h_0 = \|\boldsymbol{y}_0 - \Pi_{\mathcal{C}}\boldsymbol{y}_0\|$.

We first show that the mode estimate sequence $\{\boldsymbol{y}_t\}$ satisfies $\boldsymbol{y}_t \in S_{h_t}$ for all t under the assumptions of the corollary. One has $\boldsymbol{y}_0 \in S_{h_0}$ by the definition of h_0. Assume $\boldsymbol{x} \in S_{h_t}$ and $\boldsymbol{m}(\boldsymbol{x}) \neq \boldsymbol{0}$, and let $\boldsymbol{y} = \boldsymbol{x} + \epsilon\mathbf{U}\boldsymbol{m}(\boldsymbol{x})$ with $\epsilon \in [\gamma, 2-\gamma]$ and $\mathbf{U}$ satisfying $\mathbf{U}\boldsymbol{m}(\boldsymbol{x}) \neq \boldsymbol{0}$ and (37). Since $\mathbf{U}$ is a projection operator, one has $\mathbf{U}\boldsymbol{m}(\boldsymbol{x}) \cdot \boldsymbol{m}(\boldsymbol{x}) = \|\mathbf{U}\boldsymbol{m}(\boldsymbol{x})\|^2$. From (37), the inequality $\|\mathbf{U}\boldsymbol{m}(\boldsymbol{x})\| \geq \delta\|\boldsymbol{m}(\boldsymbol{x})\|$ holds. From Lemma 1, one has $\boldsymbol{x} + \boldsymbol{m}(\boldsymbol{x}) \in \mathcal{C}$, and thus $\|\boldsymbol{y} - \Pi_{\mathcal{C}}\boldsymbol{y}\| \leq \|\boldsymbol{y} - \boldsymbol{x} - \boldsymbol{m}(\boldsymbol{x})\|$. The right-hand side is further bounded as

$$\begin{aligned}\|\boldsymbol{y} - \boldsymbol{x} - \boldsymbol{m}(\boldsymbol{x})\|^2 &= \|\boldsymbol{m}(\boldsymbol{x})\|^2 - \epsilon(2-\epsilon)\|\mathbf{U}\boldsymbol{m}(\boldsymbol{x})\|^2 \\ &\leq \alpha^2\|\boldsymbol{m}(\boldsymbol{x})\|^2,\end{aligned} \tag{41}$$

yielding

$$\begin{aligned}\|\boldsymbol{y} - \Pi_{\mathcal{C}}\boldsymbol{y}\| &\leq \alpha\|\boldsymbol{m}(\boldsymbol{x})\| \\ &\leq \alpha(\|\boldsymbol{x} + \boldsymbol{m}(\boldsymbol{x}) - \Pi_{\mathcal{C}}\boldsymbol{x}\| + \|\Pi_{\mathcal{C}}\boldsymbol{x} - \boldsymbol{x}\|) \\ &\leq \alpha(D + h_t) = h_{t+1},\end{aligned} \tag{42}$$

where the second inequality is due to the triangle inequality. The above formula shows that if $\boldsymbol{y}_t \in S_{h_t}$, then $\boldsymbol{y}_{t+1} = \boldsymbol{y}_t + \epsilon_t\mathbf{U}_t\boldsymbol{m}(\boldsymbol{y}_t) \in S_{h_{t+1}}$ holds, thereby proving that the mode estimate sequence $\{\boldsymbol{y}_t\}$ generated by the extended MS algorithm $\boldsymbol{y}_{t+1} = \boldsymbol{y}_t + \epsilon_t\mathbf{U}_t\boldsymbol{m}(\boldsymbol{y}_t)$ with $\epsilon_t \in [\gamma, 2-\gamma]$ satisfies $\boldsymbol{y}_t \in S_{h_t}$ for all t. Since $h_t = h^* + \alpha^t(h_0 - h^*)$ with $h^* = D\alpha/(1-\alpha)$, the sequence $\{h_t\}$ converges geometrically to h^*. It also implies that for an arbitrary $h' > h^*$, there exists T such that for all $t > T$ one has $\boldsymbol{y}_t \in S_{h'}$.

Fix $h' > h^*$ and consider $\sum_{i=1}^n w_i G(\boldsymbol{y} - \boldsymbol{x}_i)$ as a function of $\boldsymbol{y}$. Since $S_{h'}$ is a bounded and closed set, the minimum of the function over $S_{h'}$ exists, and we let $\phi > 0$ denote the minimum.

In order to prove the corollary, we make use of the inequality

$$\begin{aligned}&\hat{p}(\boldsymbol{y}_{t+1}) - \hat{p}(\boldsymbol{y}_t) \\ &\geq \frac{1}{2}(\|\boldsymbol{m}(\boldsymbol{y}_t)\|^2 - \|\boldsymbol{y}_{t+1} - \boldsymbol{y}_t - \boldsymbol{m}(\boldsymbol{y}_t)\|^2)\sum_{i=1}^n w_i G(\boldsymbol{y}_t - \boldsymbol{x}_i),\end{aligned} \tag{43}$$

which is derived from (14). The above argument shows that

$$\sum_{i=1}^{n} w_i G(\boldsymbol{y}_t - \boldsymbol{x}_i) \geq \phi \tag{44}$$

holds for all $t > T$. Also, from (41) one has

$$\|\boldsymbol{m}(\boldsymbol{y}_t)\|^2 - \|\boldsymbol{y}_{t+1} - \boldsymbol{y}_t - \boldsymbol{m}(\boldsymbol{y}_t)\|^2 \geq (1-\alpha^2)\|\boldsymbol{m}(\boldsymbol{y}_t)\|^2. \tag{45}$$

Collecting the above inequalities, we have shown that the inequality

$$\hat{p}(\boldsymbol{y}_{t+1}) - \hat{p}(\boldsymbol{y}_t) \geq \frac{1-\alpha^2}{2}\phi\|\boldsymbol{m}(\boldsymbol{y}_t)\|^2 \tag{46}$$

holds for all $t > T$. Since Theorem 1 has shown that the left-hand side of (46) converges to 0 as $t \to \infty$, we have proved that $\lim_{t\to\infty} \boldsymbol{m}(\boldsymbol{y}_t) = \boldsymbol{0}$ holds.

Moreover, if G is bounded, $\sum_{i=1}^{n} w_i G(\boldsymbol{y} - \boldsymbol{x}_i)$ as a function of $\boldsymbol{y}$ takes a maximum $\psi < \infty$ in $S_{h'}$. It then follows that $\eta_t = (\sum_{i=1}^{n} w_i G(\boldsymbol{y}_t - \boldsymbol{x}_i))^{-1}$ satisfies $0 < 1/\psi \leq \eta_t \leq 1/\phi < \infty$ for all $t > T$. Convergence of $\{\nabla\hat{p}(\boldsymbol{y}_t)\}$ then follows from that of $\{\boldsymbol{m}(\boldsymbol{y}_t)\}$ and the relation $\boldsymbol{m}(\boldsymbol{y}_t) = \eta_t \nabla\hat{p}(\boldsymbol{y}_t)$. □

It should be noted, however, that convergence of the density estimate sequence $\{\hat{p}(\boldsymbol{y}_t)\}$ stated in Theorem 1 does not generally imply convergence of the mode estimate sequence $\{\boldsymbol{y}_t\}$. As described in Section 1, there is thus far no rigorous proof of convergence of mode estimate sequence $\{\boldsymbol{y}_t\}$ in the multi-dimensional cases. In the following theorem, we provide a convergence proof of the MS-type algorithms with an analytic kernel function, which is based on the convergence theories of the gradient ascent algorithm for an analytic function [35].

Theorem 2. *Under the assumptions of Theorem 1, assume further that K is an analytic function, that $\epsilon_t \in (0, 2-\gamma]$ for some t-independent constant $\gamma \in (0,2)$, and that $\mathbf{U}_t$ satisfies either (i) $\mathbf{U}_t\boldsymbol{m}(\boldsymbol{y}_t) = \boldsymbol{0}$ for all $t \geq T$ for some constant T or (ii) $\mathbf{U}_t\boldsymbol{m}(\boldsymbol{y}_t) \neq \boldsymbol{0}$ and*

$$\frac{\mathbf{U}_t\boldsymbol{m}(\boldsymbol{y}_t)\cdot\boldsymbol{m}(\boldsymbol{y}_t)}{\|\mathbf{U}_t\boldsymbol{m}(\boldsymbol{y}_t)\|\|\boldsymbol{m}(\boldsymbol{y}_t)\|} \geq \delta, \tag{47}$$

for any t such that $\boldsymbol{m}(\boldsymbol{y}_t) \neq \boldsymbol{0}$ and some t-independent constant $\delta \in (0,1]$. Then, the mode estimate sequence $\{\boldsymbol{y}_t\}$ converges to a single point for both cases (i) and (ii).

Note. Compared with Theorem 1 on the convergence of the density estimate sequence $\{\hat{p}(\boldsymbol{y}_t)\}$, the additional conditions in Theorem 2 on the convergence of the mode estimate sequence $\{\boldsymbol{y}_t\}$ are the analyticity of the kernel function K and the update conditions $\epsilon_t \in (0, 2-\gamma]$ and (47) of the algorithm. For example, the Gaussian kernel is an analytic function, which is covered by this theorem. The region of $\boldsymbol{y}_{t+1}$ satisfying the conditions of Theorem 2 is a subset of the improvement ball $\mathcal{I}(\boldsymbol{y}_t)$, and the former approaches the latter if choosing γ and δ sufficiently small.

Proof: The case (i) is obvious. In the following we prove the case (ii). The analyticity of $\hat{p}$ allows us to invoke Theorem 3.2 in [35] to prove either that $\lim_{t\to\infty}\|\boldsymbol{y}_t\| = \infty$ holds or that the mode estimate sequence $\{\boldsymbol{y}_t\}$ converges to a single point. In the proof of Corollary 1, we have shown that there exists T such that for all $t > T$ one has $\boldsymbol{y}_t \in S_{h'}$, which excludes the possibility of $\lim_{t\to\infty}\|\boldsymbol{y}_t\| = \infty$, thereby establishing the convergence of $\{\boldsymbol{y}_t\}$.

Now what remains is to confirm that the assumptions of Theorem 3.2 in [35] hold for $\{\boldsymbol{y}_t\}$. The assumptions, called the strong ascent conditions, consist of the primary ascent condition

$$\hat{p}(\boldsymbol{y}_{t+1}) - \hat{p}(\boldsymbol{y}_t) \geq \zeta\|\nabla\hat{p}(\boldsymbol{y}_t)\|\|\boldsymbol{y}_{t+1} - \boldsymbol{y}_t\| \tag{48}$$

for all t and for some $\zeta > 0$, and the complementary ascent condition, requiring that $\hat{p}(\boldsymbol{y}_{t+1}) = \hat{p}(\boldsymbol{y}_t)$ implies $\boldsymbol{y}_{t+1} = \boldsymbol{y}_t$.

We first confirm the primary ascent condition (48) to hold. One has, from (43),

$$\hat{p}(\boldsymbol{y}_{t+1}) - \hat{p}(\boldsymbol{y}_t) \geq \frac{\epsilon_t(2-\epsilon_t)}{2}\|\mathbf{U}_t\boldsymbol{m}(\boldsymbol{y}_t)\|^2 \sum_{i=1}^{n} w_i G(\boldsymbol{y}_t - \boldsymbol{x}_i). \tag{49}$$

Using $\|\mathbf{U}_t\boldsymbol{m}(\boldsymbol{y}_t)\| \geq \delta\|\boldsymbol{m}(\boldsymbol{y}_t)\|$ and $\nabla\hat{p}(\boldsymbol{y}_t) = \boldsymbol{m}(\boldsymbol{y}_t)\sum_{i=1}^{n} w_i G(\boldsymbol{y}_t - \boldsymbol{x}_i)$, both of which appeared in the proof of Corollary 1, the above inequality is further rewritten as

$$\begin{aligned}
&\hat{p}(\boldsymbol{y}_{t+1}) - \hat{p}(\boldsymbol{y}_t) \\
&\geq \frac{\epsilon_t(2-\epsilon_t)\delta}{2}\|\mathbf{U}_t\boldsymbol{m}(\boldsymbol{y}_t)\|\|\boldsymbol{m}(\boldsymbol{y}_t)\| \sum_{i=1}^{n} w_i G(\boldsymbol{y}_t - \boldsymbol{x}_i) \\
&\geq \frac{(2-\epsilon_t)\delta}{2}\|\boldsymbol{y}_{t+1} - \boldsymbol{y}_t\|\|\nabla\hat{p}(\boldsymbol{y}_t)\|,
\end{aligned} \tag{50}$$

confirming the primary ascent condition (48) to hold for $\{\boldsymbol{y}_t\}$ with $\zeta = \gamma\delta/2 > 0$.

The complementary ascent condition holds due to Lemma 2, since $\epsilon_t \in (0, 2-\gamma]$ assures that $\boldsymbol{y}_{t+1}$ is an interior point of $\mathcal{I}(\boldsymbol{y}_t)$. This completes the proof. □

Also, from Corollary 1, a condition to guarantee that a converged point is a stationary point is given below.

Corollary 2. *Under the assumptions of Theorem 2, assume further that $\epsilon_t \in [\gamma, 2-\gamma]$ for some t-independent constant $\gamma \in (0,1]$, and that G is bounded. Then, the mode estimate sequence $\{\boldsymbol{y}_t\}$ converges to a stationary point for the case (ii).*

Remark 4. The conditions (37) and (47) appearing in Theorem 2 and Corollary 1, 2 are called the angle condition in optimization research. Although the angle condition might not seem simple, it cannot be removed, as demonstrated in Example 4 below.

Example 4. For a two-dimensional case $\mathbb{R}^2$, let the sample set $\mathcal{D} = \{\boldsymbol{0}\}$ and the kernel K be a Gaussian kernel, for which $\boldsymbol{m}(\boldsymbol{x}) = -\boldsymbol{x}$ holds. Introduce the polar coordinate system to $\mathbb{R}^2$, and for $\boldsymbol{x} \in \mathbb{R}^2$, let $\angle\boldsymbol{x}$ denote the angular coordinate of $\boldsymbol{x}$. Given $\boldsymbol{x}$ and θ, let $\mathbf{U}_{\boldsymbol{x},\theta}$ be the projection operator onto a one-dimensional subspace such that the angular coordinate of the normal of the subspace spanned by $\mathbf{U}_{\boldsymbol{x},\theta}$ is $\angle\boldsymbol{x} + \theta$. One then has

$$\mathbf{U}_{\boldsymbol{x},\theta}\boldsymbol{m}(\boldsymbol{x}) = \sin\theta \begin{pmatrix} -\sin\theta & -\cos\theta \\ \cos\theta & -\sin\theta \end{pmatrix} \boldsymbol{x}, \tag{51}$$

and

$$\begin{aligned}
\boldsymbol{x} + \mathbf{U}_{\boldsymbol{x},\theta}\boldsymbol{m}(\boldsymbol{x}) &= \begin{pmatrix} 1-\sin^2\theta & -\sin\theta\cos\theta \\ \sin\theta\cos\theta & 1-\sin^2\theta \end{pmatrix} \boldsymbol{x} \\
&= \cos\theta\,\mathbf{T}(\theta)\boldsymbol{x},
\end{aligned} \tag{52}$$

where $\mathbf{T}(\theta)$ denotes the rotation matrix by angle θ.

Let $\{\theta_t\}$ be an angle sequence. Applying the update $\boldsymbol{y}_{t+1} = \boldsymbol{y}_t + \mathbf{U}_t \boldsymbol{m}(\boldsymbol{y}_t)$ iteratively, with $\mathbf{U}_t = \mathbf{U}_{\boldsymbol{y}_t,\theta_t}$, one has

$$\boldsymbol{y}_t = r_t \mathbf{T}(\kappa_t)\boldsymbol{y}_0, \tag{53}$$

where $r_t = \prod_{s=1}^{t} \cos\theta_s$ and $\kappa_t = \sum_{s=1}^{t} \theta_s$.

We show that when $\theta_t = 1/t$, with which the angle condition does not hold, as t tends to infinity r_t approaches a finite value, whereas κ_t diverges to infinity. The infinite product $r_\infty = \prod_{t=1}^{\infty} \cos\theta_t$ is convergent if and only if the infinite series $\sum_{t=1}^{\infty}(1-\cos\theta_t)$ converges [36, Theorem 7]. Since $1 - \theta^2/2 \le \cos\theta \le 1$ holds, one has $0 \le 1 - \cos\theta \le \theta^2/2$, and thus

$$\sum_{t=1}^{\infty}(1-\cos\theta_t) \le \frac{1}{2}\sum_{t=1}^{\infty}\theta_t^2. \tag{54}$$

For $\theta_t = 1/t$, one has $\sum_{t=1}^{\infty}\theta_t^2 = \zeta(2) = \pi^2/6$, where ζ denotes the Riemann zeta function, and thus

$$\sum_{t=1}^{\infty}(1-\cos\theta_t) \le \frac{\pi^2}{12}. \tag{55}$$

In particular, the infinite product r_∞ converges. On the other hand, it is well known that $\lim_{t\to\infty}\kappa_t = \infty$ for $\theta_t = 1/t$. These results imply that the forward limit set of the sequence $\{\boldsymbol{y}_t\}$ is the circle of radius $r_\infty > 0$ centered at the origin. Therefore, the sequence $\{\boldsymbol{y}_t\}$ does not converge. □

Even if analyticity of the kernel is not assumed, convergence of the mode estimate sequence in a one-dimensional problem can be established as follows.

Proposition 4. *Consider a one-dimensional case. Assume that K has a bounded, convex, differentiable, and strictly decreasing profile and G is bounded and strictly decreasing. Then the mode estimate sequence $\{y_t\}$ obtained via the iteration $y_{t+1} = y_t + \epsilon_t m(y_t)$, $\epsilon_t \in (0,1]$, converges.*

The uniform-weight case $w_i = 1/n$, $i = 1,\dots,n$, of this proposition has been proved in [23, Theorem 1]. One can prove Proposition 4 along the same way as in [23], via showing that for a sufficiently large T the mode estimate sequence $\{y_t\}_{t>T}$ is monotonic and hence converges. We omit the proof of Proposition 4, however, since the proof in [23] itself is quite involved and the extension to the weighted case is straightforward. It should be noted that the case $\epsilon_t > 1$ cannot be proved via the same strategy, since the monotonicity of the mode estimate sequence no longer holds.

Remark 5. Proposition 4 does not guarantee that the mode estimate sequence including the initial point is monotonic. Even if it converges to a mode, that point is not necessarily the mode (or stationary point) nearest to the initial point, as demonstrated in Example 5 below. This is incorrectly claimed in several papers [3], [12], [23].

Example 5. Let the sample set $\mathcal{D} = \{-0.5, 0.5\}$, the weight set $\mathcal{W} = \{1/3, 2/3\}$, and the initial mode estimate $y_0 = -1.5$ or -5. We estimated modes of the KDE $\hat{p}$ using $K(x) \propto (1+x^2/0.1)^{-1}$. The kernel function $K(x)$ is the PDF of Cauchy distribution (scaled by $\sqrt{0.1}$), and has a bounded, convex, differentiable, and strictly decreasing profile. We repeated the iteration of the conventional MS algorithm.

The KDE $\hat{p}$, the mode estimate path (34), and the density estimate path (35) associated with $\{y_t\}_{t=0,\dots,5}$ are shown in Fig. 4. When $y_0 = -1.5$, the mode estimate sequence converges to the mode nearest to the initial points, but the sequence is not monotonic. In the case where $y_0 = -5$, the mode estimate sequence is monotonic, but the sequence converges to the mode farther from the initial point. In both cases, it is commonly observed that the density estimate path $D(\tau)$ is not monotonically increasing. □

We next show that, when a Gaussian kernel is used, one can obtain not only ascent property of density estimate sequences but also that of density estimate paths.

Theorem 3. *Assume that a kernel function K is a Gaussian kernel. Let the mode estimate sequence $\{\boldsymbol{y}_t\}$ be obtained via the iteration $\boldsymbol{y}_{t+1} = \boldsymbol{y}_t + \epsilon_t \mathbf{U}_t \boldsymbol{m}(\boldsymbol{y}_t)$, $\epsilon_t \in (0,1]$, where $\mathbf{U}_t$ is a symmetric projection operator. Then, the density estimate path $D(\tau)$ associated with the sequence $\{\boldsymbol{y}_t\}$ is non-decreasing with respect to τ.*

Proof: If $\mathbf{U}_t\boldsymbol{m}(\boldsymbol{y}_t) = \mathbf{0}$, then one has $\boldsymbol{y}_{t+1} = \boldsymbol{y}_y$ and thus $D(\tau)$ takes a constant value for $\tau \in [t, t+1]$. Assume $\mathbf{U}_t\boldsymbol{m}(\boldsymbol{y}_t) \neq \mathbf{0}$. Then, one has

$$\begin{aligned}\left\|\boldsymbol{y}_{t+1} - \boldsymbol{y}_t - \frac{1}{2}\boldsymbol{m}(\boldsymbol{y}_t)\right\|^2 &= \left\|\epsilon_t \mathbf{U}_t\boldsymbol{m}(\boldsymbol{y}_t) - \frac{1}{2}\boldsymbol{m}(\boldsymbol{y}_t)\right\|^2 \\ &= (\epsilon_t^2 - \epsilon_t)\|\mathbf{U}_t\boldsymbol{m}(\boldsymbol{y}_t)\|^2 + \left\|\frac{1}{2}\boldsymbol{m}(\boldsymbol{y}_t)\right\|^2 \le \left\|\frac{1}{2}\boldsymbol{m}(\boldsymbol{y}_t)\right\|^2,\end{aligned} \tag{56}$$

which implies that $\boldsymbol{y}_{t+1}$ is in the ascent ball $\mathcal{A}(\boldsymbol{y}_t)$. From Lemma 3, one has shown that the density estimate path $D(\tau)$ associated with $\{\boldsymbol{y}_t\}$ is non-decreasing. □

Remark 6. In the cases where different bandwidth values are used for different appearances of the kernel function, Theorem 3 may not hold even if a Gaussian kernel is used, as Example 6 below demonstrates.

Example 6. Let the sample set $\mathcal{D} = \{-0.5, 0.5\}$, the weight set $\mathcal{W} = \{1/4, 3/4\}$, and the initial mode estimate $y_0 = -1.5$ or -5. We confirm properties of the KDE using different values of bandwidth in the two appearances of the kernel function and the corresponding MS algorithm with the following setting:

$$\hat{p}(x) = \sum_{i=1,2} w_i K_i(x - x_i), \tag{57}$$

$$m(x) = \frac{\sum_{i=1,2} x_i w_i G_i(x - x_i)}{\sum_{i=1,2} w_i G_i(x - x_i)} - x, \tag{58}$$

where $K_i(x) = (2\pi\sigma_i^2)^{-1/2} e^{-x^2/(2\sigma_i^2)}$, $G_i(x) = \sigma_i^{-2} K_i(x)$, $\sigma_1 = 0.2$ and $\sigma_2 = 0.4$. We repeated the iteration $y_{t+1} = y_t + m(y_t)$ using the MS function (58) for 5 times to estimate modes of the KDE (57).

The KDE (57), the mode estimate path (34), and the density estimate path (35) associated with $\{y_t\}_{t=0,\dots,5}$ are shown in Fig. 5. In every case, it is commonly observed that the density estimate path $D(\tau)$ decreases on part of the interval $(0,1)$. For this reason, when $y_0 = -1.5$, the mode estimate sequence including the initial point is not monotonic, and in the other case, the mode estimate sequence converges to the mode farther from the initial point. □

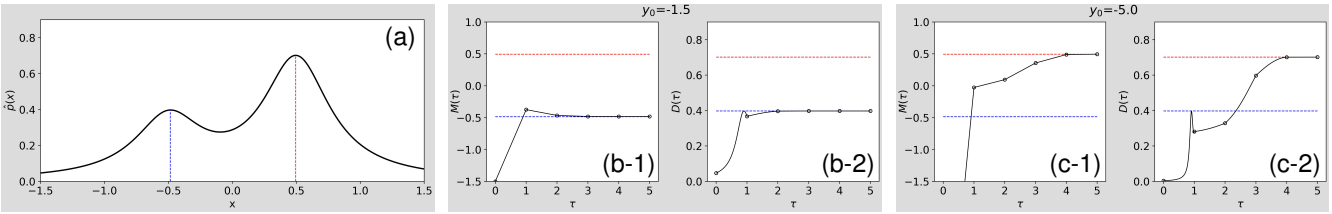

Fig. 4. KDE and mode/density estimate paths for the experiment in Example 5. The red and blue dotted lines represent the locations of the modes near $x = 0.5$ and -0.5, respectively, and the density estimates at these modes. (a) shows the KDE. (b-1) and (c-1) show the mode estimate paths $M(\tau)$ in solid lines and the mode estimate sequences $\{y_t\}_{t=0,\dots,5}$ as points. (b-2) and (c-2) show the density estimate paths $D(\tau)$ in solid lines and the density estimate sequences $\{\hat{p}(y_t)\}_{t=0,\dots,5}$ as points. {(b-1),(b-2)} and {(c-1),(c-2)} are the results with $y_0 = -1.5, -5$, respectively.

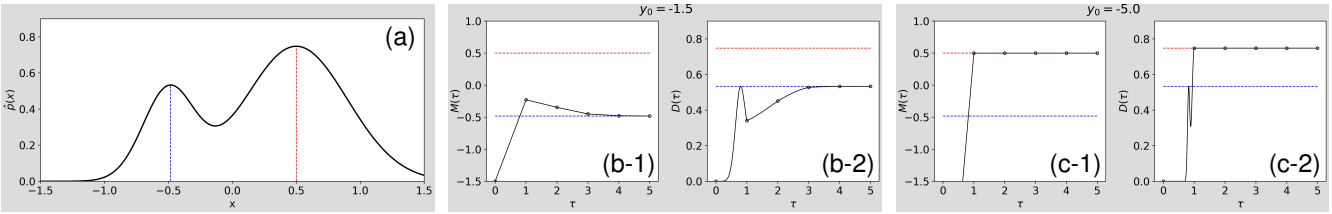

Fig. 5. KDE and mode/density estimate paths for the experiment in Example 6. The red and blue dotted lines represent the locations of the modes near $x = 0.5$ and -0.5, respectively, and the density estimates at these modes. (a) shows the KDE. (b-1) and (c-1) show the mode estimate paths $M(\tau)$ in solid lines and the mode estimate sequences $\{y_t\}_{t=0,\dots,5}$ as points. (b-2) and (c-2) show the density estimate paths $D(\tau)$ in solid lines and the density estimate sequences $\{\hat{p}(y_t)\}_{t=0,\dots,5}$ as points. {(b-1),(b-2)} and {(c-1),(c-2)} are the results with $y_0 = -1.5, -5$, respectively.

In Examples 2, 5, and 6, it was confirmed that the mode estimate sequence may jump over the mode nearest to the initial point and converge to a farther mode, depending on the choice of the step sizes and the kernel function. Below, we give a formal definition of the mode nearest to the initial point just mentioned, and discuss its properties.

Definition 6 (Proper stationary point and proper dynamical system)**.** *Given a KDE $\hat{p}$, consider the continuous-time gradient system*

$$\frac{d\boldsymbol{y}(\tau)}{d\tau} = \nabla \hat{p}(\boldsymbol{y}(\tau)), \quad \tau \in [0, \infty), \tag{59}$$

starting from an initial point $\boldsymbol{y}(0) = \boldsymbol{y}_0$. If $\boldsymbol{y}^* = \lim_{\tau\to\infty} \boldsymbol{y}(\tau)$ exists and is a stationary point of $\hat{p}$, we then define $\boldsymbol{y}^*$ as the proper stationary point *starting from $\boldsymbol{y}_0$.*

We say that a discrete-time dynamical system, generating a mode estimate sequence $\{\boldsymbol{y}_t\}$ for $\hat{p}$ given $\boldsymbol{y}_0$, is proper *if for any $\boldsymbol{y}_0$ the limit $\lim_{t\to\infty} \boldsymbol{y}_t$ of the mode estimate sequence $\{\boldsymbol{y}_t\}$ equals to the proper stationary point starting from $\boldsymbol{y}_0$, whenever the latter exists.*

This definition of the term *proper* is related to the standard definition of a cluster in mode clustering [5], [34] as the basin of attraction of the limit point $\boldsymbol{y}^*$ in the gradient system (59):

$$C(\boldsymbol{y}^*) \coloneqq \{\boldsymbol{y}_0 \in \mathbb{R}^d : \boldsymbol{y}(0) = \boldsymbol{y}_0, \lim_{\tau\to\infty} \boldsymbol{y}(\tau) = \boldsymbol{y}^*\}. \tag{60}$$

Therefore, if a discrete-time dynamical system is proper for any $\hat{p}$, then the system is able to perform mode clustering properly.

The following theorem guarantees that, under the conditions that the problem is one-dimensional and a Gaussian kernel is used, the MS algorithm $y_{t+1} = y_t + m(y_t)$ is proper.

Theorem 4. *Consider a one-dimensional case. Assume that a kernel function K is a Gaussian kernel. Let the mode estimate sequence $\{y_t\}$ be obtained via the iteration $y_{t+1} = y_t + \epsilon_t m(y_t)$, where $\epsilon_t \in [\gamma, 1]$ for some t-independent constant $\gamma > 0$. Then, the mode estimate sequence $\{y_t\}$ monotonically converges to the proper stationary point starting from y_0. In particular, the conventional MS algorithm is proper.*

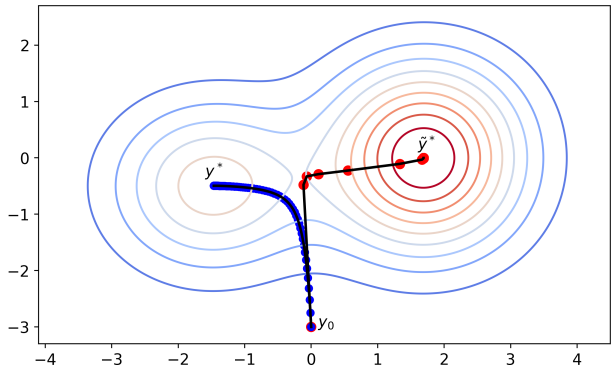

Fig. 6. KDE and mode estimate sequence for the experiment in Example 7. The KDE is plotted as contour lines: the KDE is higher for red and is lower for blue. The mode estimate sequence shown in red (resp. blue) is obtained via the iteration $\boldsymbol{y}_{t+1} = \boldsymbol{y}_t + \boldsymbol{m}(\boldsymbol{y}_t)$ (resp. $\boldsymbol{y}_{t+1} = \boldsymbol{y}_t + 0.1\boldsymbol{m}(\boldsymbol{y}_t)$) from $\boldsymbol{y}_0$, and converges to the point $\tilde{\boldsymbol{y}}^* \approx (1.69, 0.00)^T$ (resp. $\boldsymbol{y}^* \approx (-1.46, -0.50)^T$).

Proof: It has been proved in Theorem 2 that the mode estimate sequence converges to a stationary point. This, along with Theorem 1, implies that the density estimate sequence is monotonically increasing until the mode estimate sequence converges. Theorem 3 shows that the sequence monotonically converges without jumping over any stationary points. On the basis of the above three results, the proof of this theorem is completed. □

Remark 7. Even in the one-dimensional case, Theorem 4 may not hold for other kernels. Also, in the multi-dimensional case, the mode estimate sequence may not

converge to a proper stationary point even if a Gaussian kernel is used, as observed in Example 7 below.

For one-dimensional case, see Examples 2, 5, and 6 for the behaviors of mode estimate sequences when using Gaussian and other kernels.

Example 7. We give a two-dimensional example. Let the sample set $\mathcal{D} = \{(-1.5, -0.5)^T, (1.7, -0.5)^T, (1.7, 0.5)^T\}$, the weight set $\mathcal{W} = \{1/3, 1/3, 1/3\}$, and the initial mode estimate $y_0 = (0, -3)^T$. We estimated modes of the KDE $\hat{p}$ using the Gaussian kernel $K(x) = (2\pi)^{-1} e^{-\|\boldsymbol{x}\|^2/2}$.

The mode estimate sequence shown in red (resp. blue) in Fig. 6 is obtained via the iteration $\boldsymbol{y}_{t+1} = \boldsymbol{y}_t + \boldsymbol{m}(\boldsymbol{y}_t)$ (resp. $\boldsymbol{y}_{t+1} = \boldsymbol{y}_t + 0.1\boldsymbol{m}(\boldsymbol{y}_t)$) from $\boldsymbol{y}_0$ for 100 times. Then, the mode estimate sequences shown in red and blue converge to different modes from each other. In other words, the mode estimate sequence generated by the MS algorithm may not converge to the proper stationary point (mode) of $\boldsymbol{y}_0$. □

4 Acceleration Techniques of the Mean Shift Algorithm

4.1 Proposed Acceleration Techniques

On the basis of the theoretical results in Section 3, we propose two acceleration techniques of the MS-type algorithms. One is based on Theorems 1 and 2 and applicable to general cases, while the other, based on Theorem 4, is applicable when a Gaussian kernel is used and the problem is one dimensional, for example, to nonparametric modal regression with the dependent variable being one-dimensional.

As described in Remark 2, the adaptive step size (i.e., $\epsilon_t = 1$) used in the conventional MS algorithm (Algorithm 1) is supported by the fact that the lower bound of the increment $\hat{p}(\boldsymbol{y}_{t+1}) - \hat{p}(\boldsymbol{y}_t)$ given in (14) is maximized when $\epsilon_t = 1$. However, even with $1 < \epsilon_t < 2$, the algorithm $\boldsymbol{y}_{t+1} = \boldsymbol{y}_t + \epsilon_t \mathbf{U}_t \boldsymbol{m}(\boldsymbol{y}_t)$ comes with theoretical guarantees such as monotonicity of the density estimate sequence (Theorem 1) and convergence of the mode estimate sequence (Theorem 2). As shown in Example 3, the algorithm may converge faster when $1 < \epsilon_t < 2$ than when $\epsilon_t = 1$. Thus, we propose the idea of using $\epsilon_t = \epsilon \in (1, 2)$ as an acceleration method for the MS-type algorithms (Algorithm 2). This idea is based on overrelaxation[3]. Although it is also possible to adaptively determine the value of ϵ_t as in [38], it may incur extra computation, making the total computation time longer. We observed that use of a fixed value, say 1.9, as ϵ_t yields empirically good improvement in convergence speed. We note that this method can be applied even if the problem is multidimensional or using a general kernel function. The pseudo-code of Algorithm 2 is for the MS algorithm aiming to mode estimation, but the overrelaxation can also be applied to the CMS for nonparametric mode regression and the SCMS for principal curve estimation. It should however be noted that using $\epsilon_t > 1$ in mode clustering, in which the destination of the mode estimate sequence is important, can degrade the clustering accuracy.

The second acceleration method proposed in this paper is based on Theorem 4. Even in the one-dimensional case, current implementations of the MS algorithm use the multi-start method, which often uses mesh points in the data domain as initial points, as shown in Algorithm 1 [11], [12], [23]. In such implementations, it is necessary to use a sufficient number of initial points so as not to overlook any mode. Consequently, the same region may be searched many times, making the algorithm inefficient. Also, since mode estimate results are generated by the number of initial points, it requires post-processing such as grouping mode estimate results close to each other. This problem occurs in Algorithm 2 as well. Thus, we propose Algorithm 3, in which the mode is searched in one direction from the bottom to the top of the data domain. This eliminates searching the same area multiple times and the post-processing, so that it operates more efficiently.

3. A well-known and well-established example of overrelaxation is the successive overrelaxation (SOR) [37] for accelerating the Gauss-Seidel method for solving linear equations.

Algorithm 1 & **2** Conventional MS (Algorithm 1) and Over-relaxed MS (Algorithm 2) for mode estimation

Input: Sample set $\{\boldsymbol{x}_i \in \mathbb{R}^d\}_{i=1}^n$, weight set $\{w_i > 0\}_{i=1}^n$, kernel function K, initial points $\{\boldsymbol{m}_i \in \mathbb{R}^d\}_{i=1}^l$, conventional step size $\epsilon = 1$ in Algorithm 1 or acceleration parameter $\epsilon \in (1, 2)$ in Algorithm 2, and threshold δ

$\mathcal{M} \leftarrow \emptyset$
for $i = 1, \ldots, l$ **do**
  $\boldsymbol{y}_0 \leftarrow \boldsymbol{m}_i$
  repeat $\boldsymbol{y}_{t+1} \leftarrow \boldsymbol{y}_t + \epsilon \boldsymbol{m}(\boldsymbol{y}_t)$ **until** $|\boldsymbol{m}(\boldsymbol{y}_t)| < \delta$
  Add $\boldsymbol{y}_{t+1}$ to $\mathcal{M}$
end for

Output: Estimated mode set $\mathcal{M}$

Algorithm 3 One-way search acceleration of MS for mode estimation in $\mathbb{R}$

Input: Sample set $\{x_i \in \mathbb{R}\}_{i=1}^n$, weight set $\{w_i > 0\}_{i=1}^n$, Gaussian kernel K, skip size s, and threshold δ

$\mathcal{M} \leftarrow \emptyset$
$y_0 \leftarrow \min\{x_i\}$
repeat
  if $m(y_0) \geq 0$ **then**
    repeat $y_{t+1} \leftarrow y_t + m(y_t)$ **until** $m(y_t) < \delta$
    Add y_{t+1} to $\mathcal{M}$, and $y_0 = y_{t+1} + s$
  else $\{m(y_0) < 0\}$
    $y_0 = y_0 + s$
  end if
until $y_0 > \max\{x_i\}$

Output: Estimated mode set $\mathcal{M}$

4.2 Numerical Experiment

In this subsection, we compare the efficiency of the baseline algorithm using the conventional step size and multi-start method (Algorithm 1), the overrelaxation-based acceleration (Algorithm 2), and the one-way search acceleration (Algorithm 3) for mode estimation in the one-dimensional case. For the purpose, we performed a numerical experiment using real-world data.

We used the dataset "Individual household electric power consumption Data Set" [39], which includes measurements of electric power consumption in one household

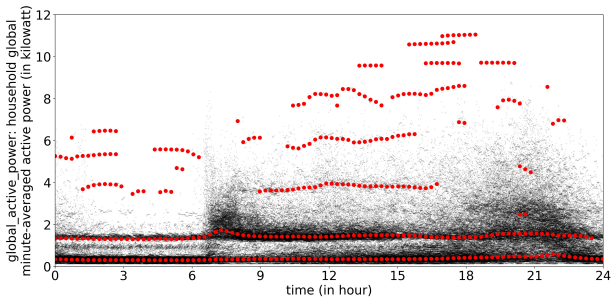

Fig. 7. Sample points and conditional mode estimates obtained by Algorithm 1 with $l = 30$ for the numerical experiment in Subsection 4.2. The black dots represent the sample points and red circles represent the mode estimates of the conditional PDF.

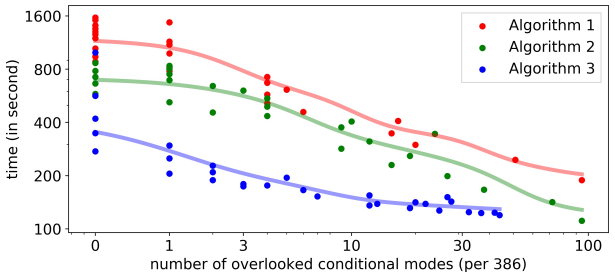

Fig. 8. The number of the overlooked conditional modes (per 386) and the total execution time (in second) by Algorithms 1, 2, and 3 for the numerical experiment in Subsection 4.2. The vertical axis is log-scaled and the horizontal axis is scaled by the so-called log-plus-one transformation $\log(\cdot + 1)$.

with a one-minute sampling rate for 47 months from December 2006 to November 2010. Although this dataset contains $365 \times 24 \times 60 = 525\,600$ sample points in the year 2009, we used $n = 521\,320$ points excluding missing values. Let the independent variable be the time of a day in hours, $\{t_i \in [0, 24)\}_{i=1}^n$, and the dependent variable be the third attribute of this dataset, that is, "global_active_power: household global minute-averaged active power (in kilowatt)", $\{x_i \in \mathbb{R}\}_{i=1}^n$. We then analysed the relationship between the independent variable and the dependent variable, using nonparametric modal regression. For 100 points t which divide $[0, 24]$ at equal intervals, we estimated modes of the conditional PDF of the dependent variable conditioned on the time t, using the CMS algorithm. The common settings for all of Algorithms 1, 2, and 3 were that the weights were $w_i = e^{-(t-t_i)^2/(2\sigma_t^2)}/(\sum_{j=1}^n e^{-(t-t_j)^2/(2\sigma_t^2)})$, $i = 1, \ldots, n$, that the kernel function was the Gaussian kernel $K(x) = (2\pi\sigma_x^2)^{-1/2} e^{-x^2/(2\sigma_x^2)}$, and that the threshold was $\delta = 10^{-6}$, where $\sigma_t = \sqrt{0.75}$ and $\sigma_x = \sqrt{0.75}$. In this situation, the total number of underlying conditional modes is 386. For Algorithms 1 and 2, the initial points were set as $m_i = \min\{x_j\} + (i-1)(\max\{x_j\} - \min\{x_j\})/(l-1)$, $i = 1, \ldots, l$, where we examined the cases with $l = 4, 5, \ldots, 30$, except for the cases $l = 1, 2, 3$ where algorithms certainly overlook modes. We set acceleration parameter ϵ of Algorithm 2 to 1.9. For Algorithm 3, the skip size was set to $s = 0.01, 0.02, \ldots, 0.3$.

The results of this experiment are summarized in Figs. 7 and 8. For all the three algorithms there is trade-off between the total execution time and the number of overlooked modes: The number of overlooked modes can be reduced by increasing the number l of the initial points in Algorithms 1 and 2, and by employing a smaller skip size s in Algorithm 3, at the price of increase in the total execution time. The results in Fig. 8 show that Algorithm 2 achieves better trade-off than Algorithm 1, and that Algorithm 3 achieves even better trade-off than the other two.

5 Conclusions

In this paper, we have provided several new results on properties and convergence of the MS-type algorithms, by theoretically studying properties of the MS function, behaviors of mode estimate sequences generated by the MS-type algorithms, and the corresponding density estimate sequences (see Table 1). Novel notions of the improvement/ascent/Lipschitz balls have been introduced, and properties of KDE $\hat{p}$, as well as its gradient, have been elucidated in terms of these notions. These properties have then been used to prove several properties of the MS-type algorithms, including Theorem 1, which shows that the density estimate sequence is non-decreasing and converges even using a step size of up to twice as large as that used in the conventional MS algorithms, Theorem 2, which shows that mode estimate sequences generated by the MS-type algorithms with a kernel function which is analytic converge under general settings, and Theorem 3, which shows monotonic increase of the density estimate path generated by the MS-type algorithms with a Gaussian kernel. We have also shown in Theorem 4 that in the one-dimensional case the mode estimate sequence generated by using a Gaussian kernel monotonically converges to the nearest stationary point.

On the basis of these results, we have proposed two acceleration techniques of the MS algorithm. One of them is based on the idea of overrelaxation and is applicable to general settings. The other one is applicable when the problem is one-dimensional and a Gaussian kernel is used. The efficiency of both proposed techniques was confirmed via the numerical experiment.

Appendix

Lemma 4. *Let $C \subset \mathbb{R}^d$ be a closed convex set and let $\boldsymbol{y} \in \mathbb{R}^d \backslash C$. Also, let $\boldsymbol{x}_0 \in C$ be the point in C that is the closest to $\boldsymbol{y}$. Then $(\boldsymbol{y} - \boldsymbol{x}_0) \cdot (\boldsymbol{y} - \boldsymbol{x}) > 0$ holds for any $\boldsymbol{x} \in C$.*

Proof: We will show that for any $\boldsymbol{x} \in C$ the inequality

$$(\boldsymbol{y} - \boldsymbol{x}_0) \cdot (\boldsymbol{x}_0 - \boldsymbol{x}) \geq 0 \tag{61}$$

holds. The proof of the lemma will be immediate by adding $\|\boldsymbol{y} - \boldsymbol{x}_0\|^2 > 0$ to both sides of (61).

We prove (61) by contradiction. Assume that there exists $\boldsymbol{x} \in C$ for which $(\boldsymbol{y} - \boldsymbol{x}_0) \cdot (\boldsymbol{x}_0 - \boldsymbol{x}) < 0$ holds. The three points $\boldsymbol{y}$, $\boldsymbol{x}_0$, and $\boldsymbol{x}$ defines a triangle in $\mathbb{R}^d$, and since $\boldsymbol{x}_0, \boldsymbol{x} \in C$ and C is convex, the edge $\boldsymbol{x}_0$-$\boldsymbol{x}$ is contained in C. The condition $(\boldsymbol{y} - \boldsymbol{x}_0) \cdot (\boldsymbol{x}_0 - \boldsymbol{x}) < 0$ implies that the angle of the triangle at the vertex $\boldsymbol{x}_0$ is less than the right angle. It in turn implies that on the edge $\boldsymbol{x}_0$-$\boldsymbol{x}$ there is a point in C closer to $\boldsymbol{y}$ than $\boldsymbol{x}_0$. This, however, contradicts the assumption that $\boldsymbol{x}_0$ is the point in C that is the closest to $\boldsymbol{y}$, completing the proof. □

ACKNOWLEDGMENTS

This work was supported in part by MEXT/JSPS KAKENHI Grant Numbers JP25120008 and JP16H02878.

REFERENCES

[1] K. Fukunaga and L. Hostetler, "The estimation of the gradient of a density function, with applications in pattern recognition," *IEEE Transactions on Information Theory*, vol. 21, no. 1, pp. 32–40, 1975.
[2] Y. Cheng, "Mean shift, mode seeking, and clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 8, pp. 790–799, 1995.
[3] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603–619, 2002.
[4] K.-L. Wu and M.-S. Yang, "Mean shift-based clustering," *Pattern Recognition*, vol. 40, no. 11, pp. 3035–3052, 2007.
[5] J. E. Chacón, "Mixture model modal clustering," *Advances in Data Analysis and Classification*, vol. 12, no. 41, pp. 1–26, 2018.
[6] W. Tao, H. Jin, and Y. Zhang, "Color image segmentation based on mean shift and normalized cuts," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 37, no. 5, pp. 1382–1389, 2007.
[7] H. Guo, P. Guo, and Q. Liu, "Mean shift-based edge detection for color image," in *International Conference on Neural Networks and Brain*, vol. 2. IEEE, 2005, pp. 1118–1122.
[8] Y. Zhu, R. He, N. Xiong, P. Shi, and Z. Zhang, "Edge detection based on fast adaptive mean shift algorithm," in *International Conference on Computational Science and Engineering*, vol. 2. IEEE, 2009, pp. 1034–1039.
[9] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 5, pp. 564–577, 2003.
[10] C. Yang, R. Duraiswami, and L. Davis, "Efficient mean-shift tracking via a new similarity measure," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1. IEEE, 2005, pp. 176–183.
[11] Y.-C. Chen, C. R. Genovese, R. J. Tibshirani, and L. Wasserman, "Nonparametric modal regression," *The Annals of Statistics*, vol. 44, no. 2, pp. 489–514, 2016.
[12] J. Einbeck and G. Tutz, "Modelling beyond regression functions: an application of multimodal regression to speed–flow data," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 55, no. 4, pp. 461–475, 2006.
[13] H. Sasaki, Y. Ono, and M. Sugiyama, "Modal regression via direct log-density derivative estimation," in *International Conference on Neural Information Processing*. Springer, 2016, pp. 108–116.
[14] R. J. Hyndman, D. M. Bashtannyk, and G. K. Grunwald, "Estimating and visualizing conditional densities," *Journal of Computational and Graphical Statistics*, vol. 5, no. 4, pp. 315–336, 1996.
[15] U. Ozertem and D. Erdogmus, "Locally defined principal curves and surfaces," *Journal of Machine Learning Research*, vol. 12, pp. 1249–1286, 2011.
[16] Y. Aliyari Ghassabeh, T. Linder, and G. Takahara, "On some convergence properties of the subspace constrained mean shift," *Pattern Recognition*, vol. 46, no. 11, pp. 3140–3147, 2013.
[17] H. Sasaki, T. Kanamori, and M. Sugiyama, "Estimating density ridges by direct estimation of density-derivative-ratios," in *International Conference on Artificial Intelligence and Statistics*, vol. 54, 2017, pp. 204–212.
[18] T. Hastie and W. Stuetzle, "Principal curves," *Journal of the American Statistical Association*, vol. 84, no. 406, pp. 502–516, 1989.
[19] J. M. Saragih, S. Lucey, and J. F. Cohn, "Deformable model fitting by regularized landmark mean-shift," *International Journal of Computer Vision*, vol. 91, no. 2, pp. 200–215, 2011.
[20] Y.-C. Chen, S. Ho, P. E. Freeman, C. R. Genovese, and L. Wasserman, "Cosmic web reconstruction through density ridges: method and algorithm," *Monthly Notices of the Royal Astronomical Society*, vol. 454, no. 1, pp. 1140–1156, 2015.
[21] M. A. Carreira-Perpiñán, "Gaussian mean-shift is an EM algorithm," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 5, pp. 767–776, 2007.
[22] X. Li, Z. Hu, and F. Wu, "A note on the convergence of the mean shift," *Pattern Recognition*, vol. 40, no. 6, pp. 1756–1762, 2007.
[23] Y. Aliyari Ghassabeh, "On the convergence of the mean shift algorithm in the one-dimensional space," *Pattern Recognition Letters*, vol. 34, no. 12, pp. 1423–1427, 2013.
[24] R. A. Boyles, "On the convergence of the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 45, no. 1, pp. 47–50, 1983.
[25] E. Arias-Castro, D. Mason, and B. Pelletier, "On the estimation of the gradient lines of a density and the consistency of the mean-shift algorithm," *Journal of Machine Learning Research*, vol. 17, no. 43, pp. 1–28, 2016.
[26] C. Améndola, A. Engström, and C. Haase, "Maximum number of modes of Gaussian mixtures," 2017, arXiv preprint arXiv:1702.05066v2 [math.ST].
[27] Y. Aliyari Ghassabeh, "A sufficient condition for the convergence of the mean shift algorithm with Gaussian kernel," *Journal of Multivariate Analysis*, vol. 135, pp. 1–10, 2015.
[28] Y.-C. Chen, C. R. Genovese, and L. Wasserman, "Generalized mode and ridge estimation," 2014, arXiv preprint arXiv:1406.1803v1 [stat.ME].
[29] J. E. Chacón, T. Duong *et al.*, "Data-driven density derivative estimation, with applications to nonparametric clustering and bump hunting," *Electronic Journal of Statistics*, vol. 7, pp. 499–532, 2013.
[30] H. Zhou and X. Huang, "Bandwidth selection for nonparametric modal regression," *Communications in Statistics-Simulation and Computation*, pp. 1–17, Feb. 2018.
[31] C. Yang, R. Duraiswami, D. DeMenthon, and L. Davis, "Mean-shift analysis using quasinewton methods," in *International Conference on Image Processing*, vol. 2. IEEE, 2003, pp. II–447.
[32] M. A. Carreira-Perpiñán and C. K. I. Williams, "On the number of modes of a Gaussian mixture," in *International Conference on Scale-Space Theories in Computer Vision*. Springer, 2003, pp. 625–640.
[33] ——, "On the number of modes of a Gaussian mixture," School of Informatics, University of Edinburgh, UK, Tech. Rep. EDI-INF-RR-0159, 2003.
[34] M. Azizyan, Y.-C. Chen, A. Singh, and L. Wasserman, "Risk bounds for mode clustering," 2015, arXiv preprint arXiv:1505.00482v1 [math.ST].
[35] P.-A. Absil, R. Mahony, and B. Andrews, "Convergence of the iterates of descent methods for analytic cost functions," *SIAM Journal on Optimization*, vol. 16, no. 2, pp. 531–547, 2005.
[36] K. Knopp, *Infinite Sequences and Series*. Courier Corporation, 1956.
[37] D. Young, "Iterative methods for solving partial difference equations of elliptic type," *Transactions of the American Mathematical Society*, vol. 76, no. 1, pp. 92–111, 1954.
[38] R. Salakhutdinov and S. T. Roweis, "Adaptive overrelaxed bound optimization methods," in *International Conference on Machine Learning*, 2003, pp. 664–671.
[39] M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml

Ryoya Yamasaki received the B.E. degree from Kyoto University, Kyoto, Japan, in 2018. He is currently working toward the M.Inf. degree of Graduate School of Informatics, Kyoto University, Kyoto, Japan. His research interests are in areas of statistics and machine learning.

Toshiyuki Tanaka received the B.E., M.E., and D.E. degrees from the University of Tokyo, Tokyo, Japan, in 1988, 1990, and 1993, respectively. He is currently a professor of Graduate School of Informatics, Kyoto University, Kyoto, Japan. His research interests are in areas of information, coding, and communications theory, and statistical learning.